JAIRE ALVES FERREIRA FILHO

CARACTERIZAÇÃO DE SÍTIOS POLIMÓRFICOS E SEQUÊNCIAS REPETITIVAS, E ESTABELECIMENTO DE COLEÇÃO NUCLEAR DE CAIAUÉ

[*Elaeis oleifera* (Kunth) Cortés]

LAVRAS - MG

2015

JAIRE ALVES FERREIRA FILHO

CARACTERIZAÇÃO DE SÍTIOS POLIMÓRFICOS E SEQUÊNCIAS REPETITIVAS, E ESTABELECIMENTO DE COLEÇÃO NUCLEAR DE CAIAUÉ [*Elaeis oleifera* (Kunth) Cortés]

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Biotecnologia Vegetal, área de concentração em Biotecnologia Vegetal, para a obtenção do título de Mestre.

Orientador

Dr. Manoel Teixeira Souza Junior

Coorientadores

Dr. Eduardo Fernandes Formighieri

Dr. Alexandre Alonso Alves

LAVRAS – MG

2015

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Ferreira Filho, Jaire Alves.
Caracterização de sítios polimórficos e sequências repetitivas, e estabelecimento de coleção nuclear de caiaué [*Elaeis oleifera* (Kunth) Cortés] / Jaire Alves Ferreira Filho. –Lavras : UFLA, 2015.
112 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2015.
Orientador(a): Manoel Teixeira Souza Júnior.
Bibliografia.

1. Palma de óleo. 2. Biologia Computacional. 3. Genotipagem por Sequenciamento. 4. Conservação Genética. I. Universidade Federal de Lavras. II. Título.

JAIRE ALVES FERREIRA FILHO

CARACTERIZAÇÃO DE SÍTIOS POLIMÓRFICOS E SEQUÊNCIAS REPETITIVAS, E ESTABELECIMENTO DE COLEÇÃO NUCLEAR DE CAIAUÉ [*Elaeis oleifera* (Kunth) Cortés]

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Biotecnologia Vegetal, área de concentração em Biotecnologia Vegetal, para a obtenção do título de Mestre.

APROVADA em 26 de fevereiro de 2015.

Dr. Eduardo Fernandes Formighieri	EMBRAPA - Agroenergia
Dr. Alexandre Alonso Alves	EMBRAPA – Agroenergia
Dr. Bruno Galveas Laviola	EMBRAPA - Agroenergia

Dr. Manoel Teixeira Souza Junior
Orientador

LAVRAS – MG

2015

A todos os alunos de pós-graduação do Brasil, que corajosamente aceitam esse desafio,

OFEREÇO!

Aos meus pais, Mirtes e Jaire, as minhas irmãs, Cinthia e Tátira, pelo indispensável apoio,

DEDICO!

AGRADECIMENTOS

A DEUS pela coragem, força e determinação a mim concedidas para superar mais essa etapa da vida.

Aos meus pais, Mirtes e Jaire, que sempre acreditaram e apoiaram os meus sonhos. As minhas irmãs, Cinthia e Tátira, sempre à disposição para me ajudar em qualquer situação.

A todos os meus professores, desde a pré-escola até o mestrado, eles são os verdadeiros heróis deste país.

Ao meu orientador, Dr. Manoel Teixeira, pela orientação, apoio e oportunidade de desenvolvimento deste trabalho.

Ao Dr. Eduardo Formighieri, pela paciência, amizade, dedicação e seus ensinamentos que foram de grande relevância para a realização deste trabalho.

Ao Dr. Alexandre Alonso, pela ajuda e dedicação no processo de desenvolvimento deste trabalho.

À Dra. Tatiana de Campos, pela orientação na iniciação científica, apoio e pela amizade.

A todos os meus colegas da pós-graduação, em especial a Rayana, Natália, Luiz, Valquíria, Flávia e Luana.

A todos os meus colegas de graduação da UFAC, em especial a Ana Áurea e Rafaella Damasceno.

Aos colegas Carol, Igor e Amanda do Laboratório de Bioinformática em Bioenergia – LBB.

À Universidade Federal de Lavras e ao programa de Biotecnologia Vegetal, pela oportunidade concedida para a realização do mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudo concedida.

"O espírito sem limites é o maior tesouro do homem."

J.K. Rowling, *Harry Potter and the Deathly Hallows*.

RESUMO

Os objetivos deste estudo foram caracterizar sítios polimórficos e sequências repetitivas e estabelecer coleção nuclear em caiaué (*Elaeis oleifera*). Foi realizada uma análise de comparação genômica de um *draft* de caiaué desenvolvido pela Embrapa com 130 coberturas de *reads* da tecnologia *Illumina HiSeq* 2000 contra o *draft* de *E. oleifera* e o genoma de *E. guineensis* públicos por meio do *software* de alinhamento Nucmer. Foi feita uma busca *in silico* para identificar regiões de repetições em *tandem* e elementos transponíveis nesse *draft*. Foram mapeados um banco de sequência geradas via plataforma DArTSeq para 553 genótipos de caiaué contra o genoma de *E. guineensis* com o *software* BWA. O *software* SAMtools foi utilizado para a identificação de SNPs. Foram mapeados no genoma os modelos gênicos de Tamareira (*Phoenix dactylifera*). Para o delineamento de coleções nucleares em caiaué, foi utilizada a estratégia de maximização da diversidade (M) com 500 locos de marcadores SNPs baseados em genotipagem por sequenciamento. Foi alinhado 68,24 e 72,83% do *draft* analisado contra os genomas de *E. oleifera* e *E. guineensis*, respectivamente. Foi possível identificar 328.879 e 618.284 locos de repetições em *tandem* e de elementos transponíveis, respectivamente. Foi possível caracterizar 17.412/2.370 PAVs/SNPs e 25.203 modelos gênicos com posições únicas no genoma. Foram gerados modelos de coleção nuclear com 37, 55, 109, 127, 138, 276, 26 e 16 indivíduos. Devido ao bom ajuste dos parâmetros validados, a coleção com 109 indivíduos (20% da coleção inteira) foi a ideal para compor a coleção nuclear de caiaué. O *draft* de caiaué possui grande parte dos genomas com que foi comparado, apresentando grande parte de um genoma altamente complexo com uma tecnologia de sequenciamento de custo acessível. Mais da metade do *draft* (55%) é composto por regiões de repetição, especialmente retrotransposons, a identificação dessas repetições pode contribuir no auxílio para a montagem de uma nova versão desse *draft*. O conjunto de marcadores PAVs/SNPs mapeados proporciona uma cobertura consideravelmente homogênea ao longo do genoma e de regiões gênicas de *E. guineensis*. O modelo de coleção nuclear gerado nesse trabalho irá permitir uma melhor utilização das subamostras no melhoramento genético e conservação genética de caiaué.

Palavras-chave: Palma de óleo. Biologia computacional. Genotipagem por sequenciamento. Conservação genética.

ABSTRACT

The objectives of this study were to characterize polymorphic sites and repetitions and establish a core collection for American oil palm (*Elaeis oleifera*). The genome draft used in this study had a 130X coverage by Illumina Hiseq 2000 and was compared with the publicly available draft of *E. oleifera*, as well as with the also publicly available genome of *E. guineensis*, through Nucmer software. In silico search was made to identify regions of tandem repeats and transposable elements in this genome draft. A bank of sequences, generated by DArTSeq platform for genotypes of *E. oleifera*, was mapped against the public *E. guineensis* genome using BWA software. The SAMtools software package was used to identify SNPs. The gene models of date palm (*Phoenix dactylifera*) were mapped on the genome of American oil palm. For the design of core collections, we used the strategy of maximizing the diversity (M) with 500 loci SNPs markers based on genotyping by sequencing. 68.24 and 72.83% of the draft analyzed was aligned against the *E. oleifera* genomes and *E. guineensis*, respectively. A total of 328,879 and 618,284 of tandem repeats and transposable elements loci were identified, respectively. It was possible to characterize 17,412/2,370 PAVs/SNPs, and 25,203 gene models, with single position in the genome. Core collections models were obtained with 37, 55, 109, 127, 138, 276, 26, and 16 individuals. As a result of the optimal adjustment of the validated parameters maintained while taking the least number of accessions, the model of 109 individuals (20% of entire collection) was chosen as the ideal to establish the core collection of *E. oleifera*. The draft of *E. oleifera* generated by Embrapa sampled much of the genomes to which it was compared, representing much of this highly complex genome with an affordable cost of sequencing technology. More than half (55%) of the draft consists of repetitions, especially retrotransposons. The identification of these regions rich on repetitive sequences will contribute to adjustments in the strategy to generate to further sequence this genome. The set of PAVs/SNPs mapped markers provide a substantially uniform coverage throughout the genome and gene regions of *E. guineensis*. The core collection model generated in this study will allow an improvement of the strategy to more efficiently conserve the germoplasm of American oil palm.

Keywords: Oil palm, computational biology, genotyping by sequencing, genetic conservation.

SUMÁRIO

CAPÍTULO 1 Introdução Geral

1 INTRODUÇÃO

O Programa Nacional de Produção e Uso de Biodiesel (PNPB) foi uma iniciativa inovadora, criada em 2005, para fomentar a produção e uso desse combustível no Brasil. A produção de biocombustíveis se trata além de uma fonte alternativa de energia, também de uma oportunidade para a geração de emprego e renda no campo. Desde a constituição do PNPB, a soja tem sido a principal matéria-prima para a produção de biodiesel, no entanto, com o investimento no setor, outras oleaginosas deverão aumentar muito a sua contribuição, entre elas se destaca a palma de óleo ou dendê.

O gênero *Elaeis* é representado pelo dendê (*Elaeis guineensis* Jacq.) e caiaué (*Elaeis oleifera* Kunth). O dendê produz óleo vegetal que é utilizado na produção de biocombustível e o caiaué é utilizado na prospecção de características favoráveis no programa de melhoramento genético do dendê. Híbridos interespecíficos entre dendê e caiaué são resistentes/tolerantes ao amarelecimento fatal (AF) que não possui agente causador biótico ou abiótico conhecido.

O dendê apresenta alta produtividade de óleo (híbrido intraespecífico tenera, de 4 a 6t de óleo/ha/ano) que possui ampla utilização alimentar, cosmética e industrial. O programa de melhoramento desenvolvido no Brasil gerou, nestes últimos 20 anos, ganho de seleção nas espécies *E. guineensis* e *E. oleifera*, entre estes, o desenvolvimento de um dos três híbridos interespecíficos resistentes ao AF no mundo, denominado BRS Manicoré (CUNHA; LOPES, 2010). Os desafios atuais no avanço desse programa requerem esforços no melhoramento genético, biologia celular e genômica para as duas espécies.

A genômica tem a sua base de investigação na análise da sequência completa do DNA e, por meio do refinamento das plataformas de sequenciamento e de ferramentas de bioinformática foi possível disponibilizar um elevado número de genomas sequenciados de diversos organismos. A análise de genomas de plantas tem um adicional de complexidade devido ao tamanho da sequência e ao elevado conteúdo de repetições que dificultam o processo de montagem dos fragmentos gerados no sequenciamento. O barateamento dessas plataformas de sequenciamento tem permitido a incorporação dessa tecnologia para espécies com pouca informação genômica como é o caso de caiaué.

A Embrapa Agroenergia iniciou a construção de um banco de dados genômico de caiaué com o desenvolvimento de um *draft* do genoma desta espécie com 130 coberturas de sequenciamento, utilizando-se da tecnologia *Illumina Hiseq* 2000. Além disso, foram desenvolvidos marcadores baseados em genotipagem por sequenciamento (GBS) para 553 indivíduos de caiaué, que fazem parte do Banco Ativo de Germoplasma (BAG) desta espécie na Embrapa Amazônia Ocidental. Essa iniciativa irá propiciar ferramentas biotecnológicas que poderão ser aplicadas no melhoramento genético da espécie, como também na produção de híbridos interespecíficos.

Outra ação visando a melhor utilização do BAG de caiaué é o desenvolvimento de coleção nuclear. Coleção nuclear pode ser definida como um subconjunto de um banco de germoplasma que representa a diversidade genética da espécie de forma não redundante. A Embrapa Amazônia Ocidental possui amplos BAGs de dendê e caiaué, que são as bases para o melhoramento genético dessas espécies, que é único no país. O estabelecimento de coleção nuclear para caiaué com marcadores moleculares permitirá o refinamento desses estudos por delinear as subamostras (acessos) mais representativas desse BAG.

Os objetivos desse estudo foram caracterizar sítios polimórficos e sequências repetitivas e estabelecer coleção nuclear em caiaué (*Elaeis oleifera*).

2 REFERENCIAL TEÓRICO

2.1 Gênero *Elaeis* spp.

O gênero *Elaeis* pertence à classe Liliopsida, ordem Arecales, família Arecaceae, subfamília Arcoideae, tribo Cocoseae e subtribo Elaeidinae (DRANSFIELD et al., 2005). O gênero é formado por somente duas espécies, a palma de óleo africana ou dendê (*Elaeis guineensis* Jacq.) e a palma de óleo americana ou caiaué (*Elaeis oleifera* (Kunth) Cortés). *E. guineensis* e *E. oleifera* podem ser cruzadas entre si, com produção de híbridos interespecíficos férteis.

Morfologicamente estas duas espécies do gênero são similares. Ambas são alógamas e apresentam cariótipo 2n = 32 (CONCEIÇÃO; MÜLLER, 2000). As plantas do gênero *Elaeis* são monoicas, ou seja, as flores masculinas e femininas são produzidas na mesma planta, mas separadas em inflorescências masculina e feminina (CONCEIÇÃO; MÜLLER, 2000).

O caiaué é utilizado como fonte de características de interesse no melhoramento genético do dendê, como resistência/tolerância a doenças e boa qualidade de óleo. Endêmico da zona tropical úmida da América Latina ocorre em populações espontâneas desde o sul do México até as áreas amazônicas do Brasil e Colômbia. É comumente encontrada em áreas ribeirinhas associadas à presença humana, nas depressões ou solos íngremes de áreas de pastagem, áreas úmidas em margens de rios, tolera tanto o sombreamento quanto o alagamento (HARDON; TAN, 1969).

O dendezeiro já é cultivado comercialmente desde o início do século XX, destacando-se principalmente por ser a oleaginosa com maior produtividade de óleo. Dois tipos de óleos podem ser extraídos, o óleo de palma, extraído industrialmente da polpa do fruto e o óleo de palmiste, extraído da amêndoa (BOARI, 2008). Comercialmente, estes dois tipos de óleos são utilizados em

várias áreas da indústria. O óleo de palma possui ampla utilização na alimentação humana, na fabricação de margarinas, panificação, biscoito, massas, tortas entre outros. Já o óleo de palmiste é valorizado nas indústrias farmacêutica, cosmética e de perfumaria (BOARI, 2008).

Atualmente o óleo de dendê vem sendo foco econômico como fonte de energia renovável (URQUIAGA; ALVES; BOODEY, 2005) na forma de biocombustível. Isso ocorreu graças ao declínio das reservas naturais de petróleo e gás natural além de fatores relacionados ao aquecimento global que abriram novas perspectivas para o óleo de palma. A cultura ocupa o primeiro lugar no mercado internacional de óleos vegetais e gorduras. De acordo com o *United States Department of Agriculture* (USDA), os países com maior produção de óleo de dendê em 1000 MT (milhões de toneladas) são: Indonésia (33500), Malásia (21250), Tailândia (2250), Colômbia (1070) e Nigéria (930). O Brasil ocupa a 11º posição (340.00) (UNITED STATES DEPARTMENT OF AGRICULTURE, 2014).

2.2 Conservação de germoplasma de *Elaeis* spp.

O *Malaysian Palm Oil Board* (MPOB) possui a mais ampla e diversificada coleção de *Elaeis* spp. no mundo, resultado de inúmeras expedições e intercâmbios, e de intenso trabalho de caracterização e avaliação do germoplasma coletado. A partir da avaliação e seleção de sua coleção, o MPOB obteve genótipos com diversas características tradicionalmente utilizadas como critério de seleção, como rendimento em óleo e altura de planta, além do desenvolvimento de cultivares direcionadas a nichos de mercado, por exemplo, qualidade de óleo (KUSHAIRI; RAJANAIDU, 2000).

O banco de germoplasma de dendê da Embrapa Amazônia Ocidental ocupa aproximadamente 25 ha, com um total de 330 subamostras (acessos)

oriundas de inúmeras regiões de sete países da África (50 subamostras da coleção do Cirad + 34 subamostras do Porim) e palmeiras subespontâneas da Bahia (246 subamostras) (RIOS et al., 2012). Já a coleção de caiaué ocupa aproximadamente 29 ha, representada por 238 subamostras coletadas em 53 locais de coleta diferentes na Amazônia Brasileira, ao longo dos rios Solimões, Negro e Madeira, na região de Manaus, no Estado do Amazonas, e ao longo do eixo rodoviário Manaus-Boa Vista, no Estado de Roraima, representando 17 populações distintas. A coleção do Rio Madeira, origem de Manicoré, é a melhor representada, devido à qualidade das plantas e aos bons índices de germinação (RIOS et al., 2012).

A Embrapa iniciou, na década de 1980, um programa de melhoramento visando ao desenvolvimento de híbridos interespecíficos (HIE) entre o caiaué e o dendezeiro. Inicialmente foram instalados experimentos para avaliar a capacidade de combinação entre diferentes origens de caiaué e de dendezeiro africano e para avaliar também a produção e o crescimento das plantas. Por meio desse programa de melhoramento genético da palma de óleo, foram desenvolvidos sete cultivares intraespecíficos registrados e um híbrido interespecífico (HIE).

Na América Latina, o amarelecimento fatal (AF) tem ameaçado o desenvolvimento da palmicultura, sendo responsável pela destruição de grandes plantações, e a única alternativa atual disponível para replantio em áreas afetadas pelo AF é a utilização do híbrido interespecífico, que apresenta tolerância/resistência a esse distúrbio. Essa característica de tolerância/resistência ao AF é herdada do caiaué.

No Brasil, todos os HIEs avaliados em áreas de incidência do AF, onde os plantios de dendezeiro foram totalmente dizimados por essa anomalia, demonstraram-se assintomáticos, e por isso, aceitos como resistentes/tolerantes. Os resultados dessas pesquisas indicaram que os HIEs, entre os acessos da

origem Manicoré e os africanos originados de La Mé (LM2T e LM10T), são os de melhor desempenho. Estes HIEs deram origem a cultivar BRS Manicoré (cadastrada no Registro Nacional de Cultivares – RNC sob o nº 26031), que é recomendada para o cultivo em área de incidência de AF (CUNHA; LOPES, 2010).

Outras vantagens da cultivar BRS Manicoré que cabe destaque: (i) teor de ácidos graxos insaturados mais elevado, propícia para a indústria alimentícia e a produção de biodiesel; (ii) produtividade mais elevada – até 30 toneladas/ha/ano; (iii) reduzida taxa de crescimento anual do caule – facilita a extração dos frutos, tornando a vida produtiva da planta mais longa; (iv) taxa de extração de óleo superior em cerca de 20% à de outras espécies; (v) menor suscetibilidade ao ataque de insetos desfolhantes. Por outro lado, a sua polinização é efetuada por métodos artificiais, demandando maior mão de obra.

2.3 Óleo de palma e biocombustíveis

Por possuir propriedades próximas ao diesel mineral, na atualidade, o biocombustível utilizado para o funcionamento em motores ciclo diesel é o "biodiesel" (álquil ésteres de ácidos graxos), podendo, em geral, ser usado nestes motores sem alterações em percentuais de até 20% (B20). O Plano Nacional de Produção e Uso do Biodiesel (PNPB) estabeleceu, a partir de 1/1/2008, a obrigatoriedade da mistura de biodiesel ao diesel, inicialmente em 2% do volume, hoje em 7% (B7). Em 2013 o Brasil foi o segundo maior consumidor de biodiesel, atrás somente dos Estados Unidos. Em 2014, estima-se o consumo brasileiro em 3,4 milhões de metros cúbicos, descontando-se a exportação de aproximadamente 41 mil metros cúbicos (MINISTÉRIO DAS MINAS E ENERGIA, 2015).

Dadas suas vantagens comparativas (maior disponibilidade de terras edafo-climaticamente apropriadas), existem perspectivas de que o Brasil possa, no médio prazo, se tornar um dos maiores exportadores mundiais de biodiesel. Segundo projeções de EPE (EMPRESA DE PESQUISA ENERGÉTICA, 2007), o país produzirá quase 10 milhões de toneladas de biodiesel em 2030, 12% da demanda doméstica projetada de diesel e 1/3 da demanda mundial de biodiesel prevista pela Agência Internacional de Energia para o referido ano (INTERNATIONAL ENERGY AGENCY, 2010).

Sabe-se que o custo da matéria-prima tem um grande peso sobre o preço final do biodiesel, representando entre 85% e 92% do total (EMPRESA DE PESQUISA ENERGÉTICA, 2011). Projeções da EPE de preços dos insumos graxos ao longo de 2010-2019 ressalta a competitividade do dendê frente a outras oleaginosas, em especial, a soja, cultivo cuja cadeia produtiva apresenta maior escala e grau de organização (EMPRESA DE PESQUISA ENERGÉTICA, 2011).

A ONU (Organização das Nações Unidas) reconheceu o potencial do cultivo da palma de óleo para a produção de biocombustíveis. Em 2009, foi aprovada a primeira metodologia de MDL (ACM 00073) para a produção de biodiesel a partir do plantio de oleaginosas em áreas degradadas. Essa metodologia sinaliza, via simplificações metodológicas, um uso preferencial pela palma e pinhão manso.

2.4 Genômica vegetal

O genoma é toda informação hereditária de um organismo que está codificada em seu DNA (ou, em alguns vírus, no RNA), isto inclui tanto os genes como as sequências reguladoras. *Haemophilus influenza*, foi o primeiro organismo a ter seu genoma completamente sequenciado, ainda no ano de 1995

(FLEISCHMAN et al., 1995). Em 2000, foi publicada a primeira sequência genômica de um organismo vegetal, a planta modelo *Arabidopsis thaliana* (*The Arabidopsis Genome Initiative*, 2000). Desde então, dezenas de espécies vegetais tiveram seus genomas completamente sequenciados, possibilitando um melhor entendimento da estrutura e organização desses genomas (HAMILTON; BUELL, 2012).

Embora o conteúdo genômico das angiospermas varie enormemente entre as espécies, grande parte desta variação não está relacionada à quantidade ou ao tamanho dos genes presentes no genoma (BENNETZEN; MA; DEVOS, 2005). A falta de correlação entre o tamanho do genoma e a complexidade biológica de um organismo é conhecida como paradoxo do valor C. Em vez de correlacionar com o conteúdo gênico, o tamanho do genoma frequentemente está correlacionado com a quantidade de DNA no genoma que é derivado de elementos de transposição. Os organismos com genomas grandes têm muitas sequências que se assemelham a elementos de transposição, enquanto organismos com genomas pequenos têm bem menos (GRIFFT et al., 2008). Os grandes responsáveis pela dinâmica de variação do tamanho dos genomas vegetais são os chamados elementos transponíveis (LISCH, 2011). Elementos transponíveis (TE's) são elementos móveis de DNA, que podem se mover e replicar dentro dos genomas hospedeiros, sendo amplamente prevalentes na maioria dos genomas de organismos eucariotos.

Longe de serem raros, os TE's chegam a compor mais de 50% do conteúdo total de alguns genomas. Esta relação pode ser ainda maior, passando dos 70% em alguns genomas de gramíneas (MEYERS; TINGEY; MORGANTE, 2001). Embora os principais grupos de TE's sejam ancestrais e estejam presentes em basicamente todos os reinos, esses elementos apresentam uma diversidade extrema, chegando a possuir milhares de famílias diferentes apenas no reino vegetal (MORGANTE et al., 2005). Sabe-se que ondas de

expansão e retração no número de TE's podem resultar em diferenças drásticas entre genomas sabidamente próximos (BENNETZEN; MA; DEVOS, 2005).

De acordo com o tipo de transposição, mediada por RNA ou DNA, os TE's podem ser divididos em duas classes principais. Elementos de classe I, ou retrotransposons, são os tipos mais comum de elementos genéticos móveis em genomas de eucariotos (FLAVELL et al., 1992) e se transpõem através de um intermediário de RNA por meio de um mecanismo de "copiar e colar" (GRANDBASTIEN, 1992). Assim como nos demais eucariotos, elementos de classe I são os mais abundantes em genomas vegetais (FESCHOTTE; JIANG; WESSLER, 2002). Elementos de classe II, ou transposons de DNA, utilizam um mecanismo de "cortar e colar" envolvendo a reintegração do elemento no genoma através de um intermediário de DNA (BOWEN; JORDAN, 2002). Ambas as classes se subdividem em famílias e superfamílias, dentro das quais membros compartilham semelhanças estruturais.

As grandes diferenças de tamanhos do genoma das espécies de plantas estão geralmente ligadas à presença de diferentes quantidades de retrotransposons. Geralmente, quanto maior o genoma vegetal, maior a chance de este conter uma grande quantidade de retroelementos (KUMAR; BENNETZEN, 1999). Por exemplo, genomas grandes, como o da cevada, podem chegar a ser compostos por até 70% destes elementos (VICIENT et al., 2001) enquanto em genomas pequenos, como o do arroz, estes representam apenas 17% da composição total (MCCARTHY et al., 2002).

Até recentemente existia pouca informação genômica para gênero *Elaeis* (BILLOTTE et al., 2005; LOW et al., 200O; TRANBARGER et al., 2011; UTHAIPAISANWONG et al., 2012), no entanto, o trabalho desenvolvido pelo grupo *Malaysian Palm Oil Board* (MPOB) disponibilizou um genoma de referência para *E. guineensis* e *E. oleifera* (SING et al., 2013a). Nesse trabalho foi utilizada a tecnologia de sequenciamento Roche/454 e Sanger, sendo

montado 1,5-gigabase (Gb) do genoma de *E. guineensis* (26X coberturas) e 658-megabase (Mb) ancorado ao mapa genético dessa espécie. Também foi desenvolvido um *draft* do genoma de *E. oleifera* (25X coberturas) de 458 Mb de tamanho (SING et al., 2013a).

Com base na sequência do genoma e dados de transcriptoma de *E. guineensis* foi identificado um gene específico, denominado *Shell*, que determina a natureza da casca do fruto. O tipo tenera é heterozigoto para o gene *Shell*, uma combinação que se traduz por um rendimento de óleo por fruto 30% maior que o tipo dura (homozigoto dominante), a palma de óleo tem um ciclo reprodutivo muito longo, são necessários até seis anos para que os produtores determinem o tipo de muda. A obtenção de um marcador genético permitiria acelerar o processo de seleção e reduzir a superfície cultivada, conciliando assim o aumento crescente de óleos e biocombustíveis e o plantio sustentável da espécie (SING et al., 2013b).

A Embrapa Agroenergia está envolvida em um grande projeto genômico de caiaué e dendê, com o sequenciamento *de novo* do genoma de caiaué que gerou um *draft* de 130 coberturas com a tecnologia *Illumina Hiseq 2000*, e o sequenciamento *de novo* do dendê em parceria com o consórcio internacional OPGP (*Oil Palm Genome Project*), coordenado pelo CIRAD e pela Neiker Tecnalia, respectivamente da França e Espanha. Além da geração de dados de transcriptoma para essas duas espécies.

2.5 Sequenciamento de Nova Geração (NGS)

A genômica revolucionou a maneira como a análise da informação genética de um organismo é feita, abrindo caminho para estudos que não eram concebíveis até alguns anos atrás. A tecnologia de sequenciamento do DNA passou por muitas evoluções e, aliada a grandes avanços nas ferramentas de

bioinformática, permitiu um grande salto na análise de sequências genômicas. No início do Projeto Genoma Humano, o sequenciamento de DNA era realizado manualmente, de forma que consumia muito tempo e representava intensivo trabalho, com o passar do tempo, novas tecnologias automatizadas e que produzem o número muito maior de sequências surgiram.

2.5.1 Tecnologia Roche/454 – Pirosequenciamento

O sistema 454 foi a primeira plataforma de sequenciamento de nova geração a ser comercializada. A plataforma 454 realiza o sequenciamento baseado em síntese, o pirosequenciamento (RONACHI, 2001). A leitura da sequência nesse sistema é realizada a partir de uma combinação de reações enzimáticas que se inicia com a liberação de um pirofosfato, oriundo da adição de um desoxinucleotídeo à cadeia. Em seguida esse pirofosfato é convertido para ATP, pela ATP sulfurilase, sendo este utilizado pela luciferase para oxidar a luciferina, produzindo um sinal de luz capturada por uma câmera de CCD (*charge-coupled device*) acoplada ao sistema (Figura 1).

O sistema requer que o DNA seja mecanicamente fragmentado em sequências de 300 – 800pb, transformado em fragmentos abruptos fosforilados e ligados a adaptadores de sequência específica. A biblioteca de DNA da amostra é ligada a adaptadores A e B nas extremidades 3’ e 5’ dos fragmentos, respectivamente, os quais são utilizados nas etapas posteriores de isolamento dos fragmentos (A-B) e amplificação e nas reações de sequenciamento. O adaptador B possui biotina marcado na região 5’, o que permite o isolamento dos fragmentos ligados ao adaptador A na extremidade 3’ e adaptador B na extremidade 5’ na amostra. Somente os fragmentos A e B são eluídos na reação de purificação e são especificamente ligados às microesferas que carregam várias cópias da sequência complementar exata ao adaptador B de um único

fragmento. O outro adaptador é utilizado no anelamento do *primer* que inicia a reação de sequenciamento. As microesferas ligadas aos fragmentos únicos de fita simples são então emulsionados em uma mistura de água e óleo com reagentes para amplificação clonal do fragmento fita simples em cerca de 1 milhão de cópias. Na PCR em emulsão, o óleo em solução aquosa forma micelas, nas quais as microesferas são capturadas. Cada micela funcionará como um microrreator, produzindo muitas cópias idênticas de um mesmo fragmento isoladamente em um microssuporte.

Após a PCR de emulsão, as microesferas ligadas aos fragmentos de fita simples são depositadas em poços distintos em uma placa de sílica onde os reagentes para o sequenciamento são distribuídos. As reações de sequenciamento ocorrem em cada poço, para um único tipo de fragmento ligado a microesfera, não havendo, portanto, competição por reagentes com outros fragmentos da biblioteca. A placa de sequenciamento é dividida em 1,6 milhões de poços com diâmetro suficiente para alojar uma única microesfera.

A placa de sequenciamento é inserida junto ao sistema óptico de leitura no equipamento. Os reagentes e soluções de sequenciamento são então distribuídos por toda a placa a cada ciclo para o sequenciamento paralelo dos 1,6 milhões de poços. O sequenciamento é realizado em ciclos, e a cada ciclo um tipo determinado de nucleotídeo é adicionado à reação. Se o nucleotídeo adicionado for incorporado à sequência em síntese, um sinal de luz é emitido, sendo a intensidade desse sinal um reflexo do número de nucleotídeos desse tipo específico que foram sucessivamente incorporados na molécula. Como o nucleotídeo que é adicionado a cada ciclo é conhecido, o sinal de luz pode ser diretamente utilizado como informação de sequência.

Os fragmentos sequenciados nessa plataforma passam por um sistema de análise de qualidade em que sequências distintas oriundas do sequenciamento de uma única microesfera são eliminados, bem como as leituras em que a sequência

inicial TGCA (quatro primeiros nucleotídeos dos adaptadores) não aparece. Inicialmente as leituras produzidas possuíam tamanho próximo a 100pb; porém, com a otimização do processo, estas tiveram o tamanho ampliado em aproximadamente dez vezes.

Genomas pequenos, como os de bactérias, podem ser facilmente montados usando a plataforma 454. Em relação às outras tecnologias de sequenciamento de nova geração, a plataforma gera maiores leituras, e por isso tem sido muito utilizada, inclusive para o sequenciamento de genomas eucariotos. Uma limitação importante da plataforma 454 é a baixa eficiência na determinação de homopolímeros. Como a intensidade do sinal de fluorescência relaciona-se ao número de vezes que um determinado nucleotídeo foi incorporado à sequência, a determinação precisa de sequências em que um único nucleotídeo é repetido mais de três vezes torna-se imprecisa. O custo do sequenciamento com essa plataforma é superior ao custo das plataformas Illumina, mas, nos casos em que a produção de leituras maiores são necessárias, a plataforma 454 é a melhor opção.

Embora o pirosequenciamento possua limitações particulares em termos de perfil de acurácia, a falta de uma etapa de clonagem na plataforma 454 significa que as sequências não tipicamente amostradas numa abordagem WGS (*whole-genome shotgun*), devido aos viesses de clonagem, serão mais provavelmente representados no conjunto de dados FLX, os quais contribuem para uma cobertura de genoma mais compreensiva (MEDINI et al., 2008).

2.5.2 Plataforma Illumina

O sequenciamento na plataforma Solexa, assim como o sequenciamento de Sanger, é realizado por síntese usando a DNA polimerase e nucleotídeos terminadores marcados com diferentes fluoróforos (Figura 2). A inovação dessa plataforma consiste na clonagem *in vitro* dos fragmentos em uma plataforma sólida de vidro, processo também conhecido como PCR de fase sólida (FERDUCO et al., 2006). A superfície de clonagem (*flow cells*) é dividida em oito linhas que podem ser utilizadas para o sequenciamento de até oito bibliotecas. Em cada linha, adaptadores são fixados à superfície pela extremidade 5', deixando a extremidade 3' livre para servir na iniciação da reação de sequenciamento dos fragmentos imobilizados no suporte por hibridização.

Os fragmentos de DNA da amostra são também ligados aos adaptadores em ambas às extremidades, o que permite a sua fixação ao suporte de sequenciamento por hibridização a um dos adaptadores fixados. No primeiro ciclo de amplificação, nucleotídeos não marcados são fornecidos para que haja síntese da segunda fita do fragmento imobilizado no suporte. A alta densidade de adaptadores no suporte facilita a hibridização do adaptador livre dos fragmentos imobilizados a sua sequência complementar fixa perto do clone inicial durante o ciclo de anelamento. Após o ciclo de anelamento, o fragmento forma uma estrutura em "ponte" na superfície de sequenciamento e a extensão ocorre, formando a fita complementar também em "ponte". No ciclo de desnaturação, as fitas são separadas e linearizadas. Esses ciclos são repetidos 35 vezes e assim as cerca de mil cópias geradas de cada fragmento nessa PCR de fase sólida permanecem próximas umas das outras, formando um *cluster* de sequenciamento. Etapas de desnaturação são necessárias para a separação dos duplex formados e, nos próximos ciclos de amplificação, nucleotídeos

terminadores marcados são fornecidos para as reações de sequenciamento que ocorrem dentro de cada *cluster*. A alta densidade dos *cluster* de sequenciamento possibilita que o sinal de fluorescência gerado com a incorporação de cada um dos nucleotídeos terminadores tenha uma intensidade suficiente para garantir sua detecção exata. Até 50 milhões de *clusters* podem ser produzidos por linha, correspondendo a uma representação satisfatória da biblioteca. Após a incorporação de cada nucleotídeo no fragmento em síntese, a leitura do sinal de fluorescência é realizada. Em seguida, ocorre uma etapa de lavagem para remoção dos reagentes excedentes e remoção do terminal 3' bloqueado e do fluoróforo do nucleotídeo incorporado no ciclo anterior para que a reação de sequenciamento prossiga. A leitura das bases é feita pela análise sequencial das imagens capturadas em cada ciclo de sequenciamento. Inicialmente a plataforma permitia leituras de 25-35 pb (SHENDURE; JI, 2008), atualmente com a tecnologia *HiSeq* são obtidos leituras por volta de 100 pb.

Uma das grandes desvantagens dessa plataforma é o tamanho reduzido dos *reads*, o que dificulta o processo de montagem do genoma, principalmente aqueles comgrande quantidade de repetições. No entanto, muitos genomas vegetais foram montados fazendo uso dessa tecnologia, tais como o de morango (SHULAEV et al., 2011), tomate (THE TOMATO GENOME CONSORTIUM, 2012) e trigo (LING et al., 2013).

2.6 Marcadores moleculares

Os programas de melhoramento genético de plantas têm tido grande impacto na produção agrícola, as técnicas convencionais de melhoramento foram capazes de aumentar de forma significativa a produção das principais culturas de interesse. Apesar do sucesso acumulado desses programas, ganhos adicionais na eficiência do melhoramento podem ser obtidos por meio de

aplicação de tecnologias moleculares. Nesse contexto estão os marcadores moleculares que têm se mostrado úteis em diversos aspectos no melhoramento genético de plantas.

Basicamente existem três modalidades de marcadores moleculares, aqueles baseados em PCR - Reação em Cadeia da Polimerase (MULIS; FALOONA, 1987), como o RAPD - *Random Amplified Polymorphic* DNA (WILLIAMS et al., 1990), AFLP - *Amplified Fragment Lenght Polymorphism* e o SSR - *Simple Sequence Repeat* (LITT; LUTY, 1989), aqueles baseados em hibridização, como o RFLP - *Restriction Fragment Lenght Polymorphism* (BOTSTEIN et al., 1980) e o DArT - *Diversity Arrays Technology* (JACCOUD et al., 2001), e os marcadores baseados em sequenciamento como os SNPs (*Single Nucleotide Polymorphism*) e os PAVs (*Presence/Absence Variants*).

A escolha do marcador ideal para o estudo genético e de melhoramento de plantas vai depender do objetivo do projeto e a pergunta biológica que se deseja responder. Um marcador ideal é aquele que apresenta fácil acesso e disponibilidade, rapidez de resposta e alta reprodutibilidade, permite a troca de informações entre laboratórios e entre populações e, ou, espécies diferentes, bem como automação na geração de dados e subsequente análise. Outras características desejáveis incluem as de apresentar natureza altamente polimórfica, herança codominante (identificação dos indivíduos homozigotos e heterozigotos), ocorrência frequente no genoma e seleção neutra (seleção isenta de interferência de práticas de manejo e condição ambiental).

A grande desvantagem da maioria das técnicas citadas (RAPD, AFLP, SSR e RFLP) reside na análise de pequena amostragem do genoma por ensaio, dificuldade na automação das etapas e o baixo número de locos analisados por experimento. No entanto, marcadores SSR possuem elevada informatividade devido à sua natureza codominante e multialélica, o que os tornam marcadores

ainda muito utilizados para estudos detalhados de estruturação genética, mapeamento e *fingerprint* de cultivares (TAKAYAMA et al., 2015).

Os marcadores SNPs apresentam como vantagem a distribuição e frequência no genoma e a automação dos dados, por isso são recomendados para trabalhos de seleção genômica ampla e estudos de plantas aparentadas e com base genética estreita. No entanto, esse marcador tem natureza bialélica e menor poder de resolução se comparado com os marcadores multialélicos SSR, embora essa deficiência possa ser superada pela capacidade de análise de grande número de locos por ensaio.

Os marcadores DArT, apesar de recentes, estão sendo exaustivamente testados em várias espécies. Para a maioria das espécies em que tem sido aplicada, a técnica está sendo adaptada e usada para a caracterização de coleções de germoplasma, diversidade genética, construção de mapas genéticos de alta resolução e identificação de QTLs (*Quantitative Trait Loci*). Por ser um marcador amplamente distribuído no genoma e de genotipagem em larga escala, os DArT têm um grande potencial para seleção genômica ampla. Além disso, esses marcadores têm apresentado bom desempenho nas análises de espécies poliploides.

2.7 Genotipagem por sequenciamento

Uma nova abordagem conhecida como genotipagem por sequenciamento começou a ser desenvolvida em 2010 por grupos nos Estados Unidos e pela empresa australiana DArTPty© (MYLES et al., 2010). Essa abordagem permite a automatização da genotipagem de milhares ou dezenas de marcadores, permitindo uma amostragem significativa do genoma de estudo.

A metodologia de genotipagem por sequenciamento envolve a redução da complexidade do genoma com enzimas de restrição que geram fragmentos de

diferentes tamanhos, esses fragmentos são selecionados e submetidos ao sequenciamento por meio de alguma plataforma de NGS e indexados com uma sequência adaptadora (barcodes) que permite posterior identificação da amostra. O polimorfismo é detectado pela ausência ou presença da sequência (PAVs) em diferentes indivíduos devido à variabilidade nos sítios das enzimas de restrição.

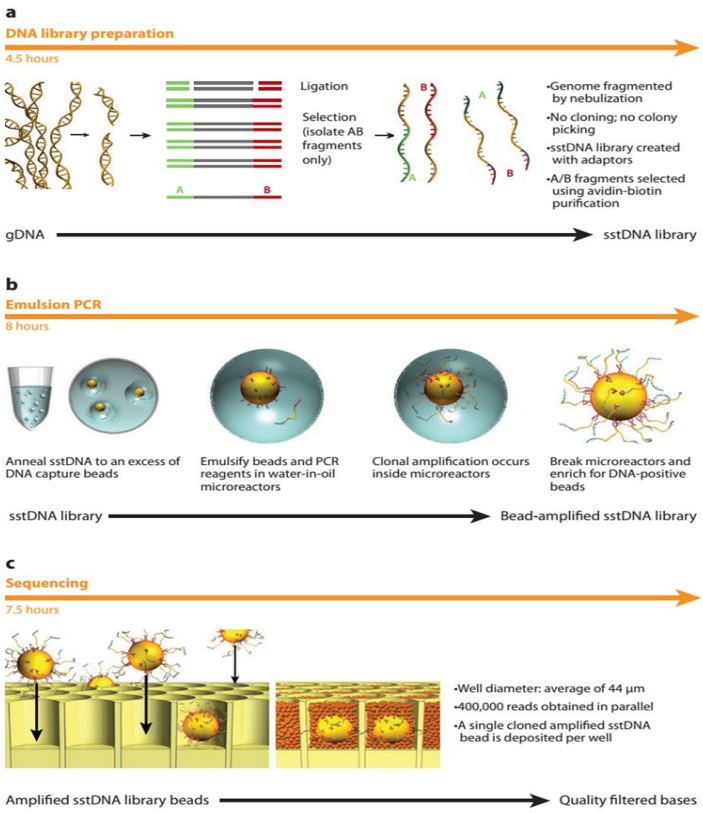

Figura 1 Método de genotipagem da plataforma de sequenciamento 454 através de PCR em emulsão

Legenda: Os fragmentos de DNA são ligados a adaptadores específicos (A, em verde, e B, em vermelho). Em seguida, os fragmentos de DNA são capturados por *beads* de agarose e isolados em micelas junto de reagentes de PCR; dentro das micelas ocorre a amplificação da sequência.

Fonte: MARDIS, 2008.

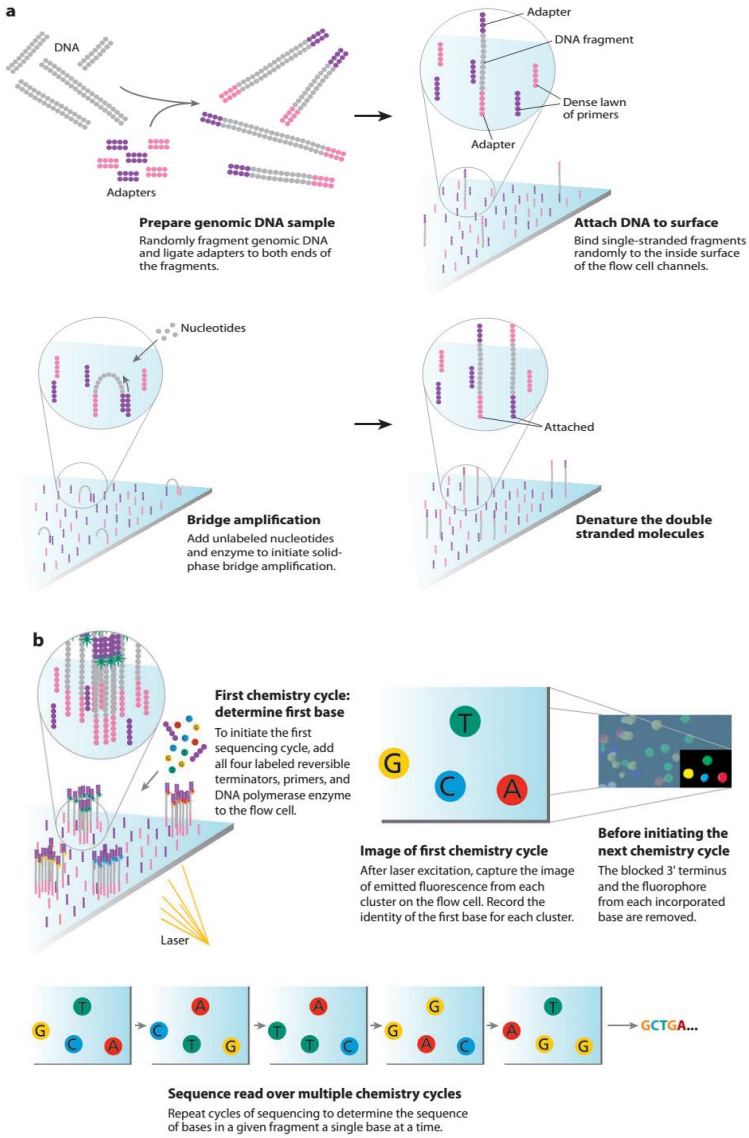

Figura 2 Metodologia de sequenciamento da plataforma Illumina, através da amplificação de fragmentos em suporte sólido

Legenda: O sequenciamento ocorre por síntese de DNA que emprega terminadores removíveis marcados fluorescentemente, adicionados separadamente à cada ciclo; a cor de cada terminador é detectada e equivale à identidade da base.

Fonte: MARDIS, 2008.

As sequências brutas obtidas (em torno de 70 pares de base) devem passar por filtros de qualidade de base. Posteriormente deve-se realizar o alinhamento das sequências contra um genoma de referência da espécie de interesse, com base em critérios de qualidade do mapeamento e de cobertura é possível realizar a chamada dos genótipos. Os marcadores PAVs geralmente possuem dados de genotipagem binários (0=ausência e 1=presença) o que os caracterizam como um marcador dominante, resultando em uma análise em que não é possível distinguir o genótipo heterozigoto. No entanto, a informatividade genética pode ser ampliada com a busca por variação de nucleotídeos dentro das sequências com a chamada de SNPs, que é um marcador codominante e bialélico no qual o heterozigoto pode ser identificado (Figura 3).

Com o barateamento das plataformas de sequenciamento de nova geração (NGS), a metodologia de genotipagem por sequenciamento (GBS) vem ganhando destaque por permitir uma cobertura bastante homogênea ao longo do genoma, ferramentas de bioinformática são de fundamental importância devido à grande quantidade de dados que são gerados.

2.8 Bioinformática e análise genômica

A bioinformática é um componente das ciências biológicas, o seu principal objetivo é desenvolver algoritmos para a resolução de problemas biológicos. Os especialistas em bioinformática utilizam e constroem ferramentas de análise de dados, sendo muito importante que conheçam os problemas biológicos, bem como as soluções computacionais, para que possam produzir soluções/ferramentas úteis.

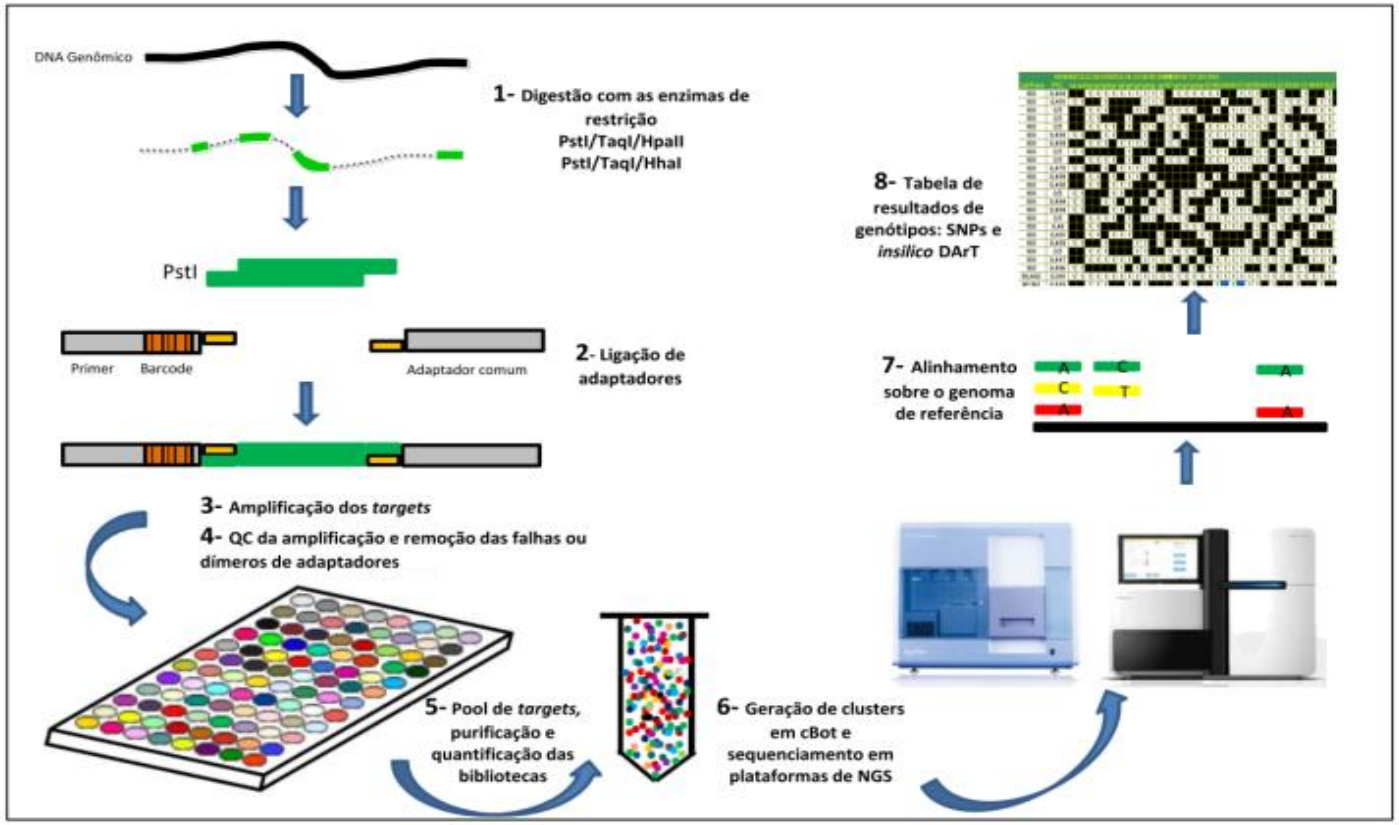

Figura 3 Fluxograma do procedimento DArT-Seq baseado em Sequenciamento de Nova Geração

Fonte: Sansaloni (2012).

O objetivo da biologia, na era dos projetos genomas, é desenvolver um entendimento quantitativo de como os seres vivos são formados com base no genoma que os codifica. Quebrar o código do genoma é uma tarefa complexa. No nível mais simples, ainda é difícil identificar os genes desconhecidos pela análise de computador da sequência genômica.

Uma longa cadeia de dados de sequência genômica é tão útil quanto um livro de referência sem cabeçalho de assunto, número de página ou índice (GIBAS; JAMBECK, 2001). Uma das principais tarefas da bioinformática é criar sistemas de *softwares* para gerenciamento de informações que podem realmente registrar cada parte de uma sequência de genoma com informações, desde a função de um gene até a estrutura do produto proteico (se houver), considerando que o gene é expresso em diferentes estágios da vida de um organismo.

Um dos maiores desafios da bioinformática na análise de genomas é a montagem dos fragmentos oriundos de sequenciamento em uma sequência contínua. A identificação das sobreposições de sequências entre fragmentos impõe algumas restrições sobre como as sequências podem ser montadas. O algoritmo de montagem tenta satisfazer todas as restrições e produzir uma ordenação ideal de todos os fragmentos que compõem o genoma. As características das sequências repetitivas podem complicar o processo de montagem. O processo de sequenciamento pode falhar em alguns casos, deixando intervalos na sequência de DNA, que devem ser resolvidos pelo ressequenciamento. Esses intervalos complicam a montagem automatizada (GIBAS; JAMBECK, 2001).

Os dados biológicos são armazenados em bancos de dados e arquivos de textos enormes. É possível classificar e analisar esses dados manualmente, mas levaria tempo demais, por isso uma linguagem de programação é de grande importância para automatizar o processo. A linguagem de programação Perl

(*Practical Extraction and Report Languange*) é utilizada para desenvolver aplicativos da *web* e é utilizada com frequência por biólogos computacionais. Os programas em Perl (denominados *scripts*) têm a extensão *.pl (ou *.cgi se forem aplicativos da web). A Perl, com sua capacidade altamente desenvolvida para detectar padrões em dados, e especialmente sequências de caracteres de texto, é a opção mais óbvia em bioinformática (SCHWARTZ; FOY; PHOENIX, 2011).

O projeto Bioperl dedica-se à criação de uma biblioteca de código aberto de módulos de pesquisa em bioinformática. A ideia geral é que itens comuns em bioinformática (como sequências e alinhamento de sequências) sejam representados como objeto em Bioperl. Atualmente, contém módulos para geração e armazenamento de alinhamento de sequências, gerenciamento de dados de anotação, análise de saída dos programas de pesquisa em banco de dados de sequências BLAST e HMMer, e há outros módulos em construção.

2.9 Coleção Nuclear

As coleções de germoplasma foram estabelecidas para preservar a diversidade genética existente, antes que muito dessa diversidade fosse perdida para sempre, devido ao uso de cultivares modernos e à exploração agrícola tecnificada. A ênfase dada na importância de preservar o germoplasma de culturas importantes tem levado à formação e manutenção de coleções muito grande. Embora a representatividade da coleção possa ser conseguida através de coleções de tamanho grande (FRANKEL; BENNETT, 1970), a acessibilidade e utilidade de uma coleção é inversamente relacionada com seu tamanho (FRANKEL; SOULÉ, 1981).

Frankel e Brown (1984) propuseram pela primeira vez o conceito de "coleção nuclear" (*core collection*) com o objetivo de melhorar a utilização e a acessibilidade ao banco. Essa coleção consistiria de um número bem menor de

acessos e representaria, com um mínimo de repetibilidade, a diversidade genética da espécie e espécies primitivas relacionadas. De acordo com Malosetti e Abadie (2001), o desenvolvimento da coleção nuclear representa uma estratégia de baixo custo para melhorar a avaliação e utilização do germoplasma e também para auxiliar as atividades do curador.

As coleções nucleares não são desenvolvidas com o objetivo de substituir a coleção de germoplasma, mas para reter grande parte da diversidade genética em uma coleção menor, de forma mais acessível aos usuários, podendo ser considerada como amostra permanente disponível, ou criada em resposta a uma necessidade específica (SPAGNOLETTI ZEULI; QUALSET, 1993). A coleção nuclear poderia ser usada para guiar mais eficientemente a utilização da coleção reserva.

Quando marcadores moleculares são utilizados para obter a coleção nuclear, na retenção da diversidade genética, métodos baseados na estratégia M (maximização) ou combinação de métodos (MARITA; RODRIGUEZ; NIENHUIS, 2000) são utilizados para selecionar combinações específicas de acessos ao mesmo tempo em que maximiza o número de alelos em cada loco. A diversidade alélica de uma coleção nuclear formada pela estratégia M é definida em termos do número de classes de alelos representados na amostra.

Geralmente em estudos de desenvolvimento de coleção nuclear são utilizados marcadores microssatélites devido ao seu elevado número de alelos por loco (TÓTH; GÁSPÁRI; JURKA, 2000), no entanto, marcadores SNPs apesar de bialélicos podem ser utilizados em larga escala, compensando quantitativamente a sua menor informatividade em comparação com microssatélites. Fato esse que levou à utilização desses marcadores na construção de coleções nucleares de diferentes espécies (BELAJ et al., 2012; OLIVEIRA et al., 2014).

3 CONCLUSÃO

Com o aumento da demanda do óleo de palma na indústria alimentícia e de biocombustíveis serão necessários esforços maiores de pesquisa e desenvolvimento no melhoramento genético de dendê e caiaué, com o objetivo de melhorar a eficiência deste processo de desenvolvimento de genótipos superiores. Esta eficiência será medida principalmente no que se refere a redução do tempo e do custo para desenvolvimento de novos genótipos desta palmeira perene.

Devido à forte redução no custo de sequenciamento e à ampla disponibilidade de plataformas de nova geração de sequenciamento é hoje possível realizar estudos de genômica em culturas que não são commodities, cujos projetos de pesquisa não contam com amplos recursos financeiros. A Embrapa vem investindo de forma intensa em estudos de genômica, genética molecular, citogenética, bioinformática e metabolômica de em *Elaeis* spp., com o objetivo de fortalecer seu programa de melhoramento genético desta palmeira, visando à obtenção de híbridos interespecíficos superiores.

REFERÊNCIAS

BELAJ, A. et al. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. **Tree Genetics & Genomes**, Amsterdam, v. 8, n. 1, p. 365-378, Dec. 2012.

BENNETZEN, J. L.; MA, J.; DEVOS, K. M. Mechanism of recent genome size variation in flowering plants. **Annals of Botany**, Oxford, v. 95, n. 1, p. 127-132, Jan. 2005.

BILLOTTE, N. et al. Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). **Theoretical and Applied Genetics**, Berlin, v. 110, n. 4, p. 754-765, Feb. 2005.

BOARI, A. de J. **Estudos realizados sobre o amarelecimento fatal do dendezeiro (*Elaeis guineensis* Jacq.).** Belém: Embrapa, 2008. 59 p. (Documento, 348).

BOTSTEIN, D. et al. Construction of a genetic linkage map using restriction fragment length polymorphisms. **The American Journal of Human Genetics**, Houston, v. 32, n. 3, p. 314-331, May 1980.

BOWEN, N. J.; JORDAN, I. K. Transposable elements and the evolution of eukaryoutic complexity. **Current Issues in Molecular Biology**, Wymondham, v. 4, n. 3, p. 65-76, July 2002.

CONCEIÇÃO, H. E. O.; MULLER, A. A. Botânica e morfologia do dendezeiro. In: VIÉGAS, I. J. M.; MULLER, A. A. (Ed.). **A cultura do dendezeiro na Amazônia brasileira.** Belém: Embrapa Amazônia Oriental, 2000. p. 31-44.

CUNHA, R. N. V.; LOPES, R. **BRS Manicoré:** híbrido interespecífico entre o caiaué e o africano recomendado para áreas de incidências de amarelecimento-fatal. Manaus: Embrapa Amazônia Ocidental, 2010. 3 p.

DRANSFIELD, J. et al.A new phylogenetic classification of the palm family, Arecaceae. **Kew Bulletin**, Richmond, v. 60, n. 4, p. 559-569, Sept. 2005.

EMPRESA DE PESQUISA ENERGÉTICA. **Plano decenal de expansão 2020**. Brasília: Ministério de Minas e Energia, 2011. 343 p.

EMPRESA DE PESQUISA ENERGÉTICA. **Plano nacional de energia 2030:** combustíveis líquidos. Brasília: Ministério de Minas e Energia, 2007. 244 p.

FERDUCO, M. et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. **Nucleic Acids Research**, London, v. 34, n. 3, p. 1-13, June 2006.

FESCHOTE, C.; JIANG, N.; WESSLER, S. R. Plant transposable elements: where genetics meets genomics. **Nature Review-Genetics**, Lodon, v. 3, n. 5, p. 329-341, May 2002.

FLAVELL, A. J. et al. Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. **Nucleic Acids Research**, London, v. 20, n. 14, p. 3639-3644, July 1992.

FLEISCHMAN, R. D. et al. Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. **Science**, Washington, v. 269, n. 5223, p. 496-512, July 1995.

FRANKEL, O. H.; BENNETT, E. **Genetic resources in plants:** their exploration and conservation. Oxford: Blackwell, 1970. 554 p.

FRANKEL, O. H.; BROWN, A. H. D. Plant genetic resources today: a critical appraisal. In: HOLDEN, J. H. W.; WILLIAMS, J. T. (Ed.). **Crop genetic resources:** conservation and evaluation. London: Harper Collins Publishers, 1984. p. 249-257.

FRANKEL, O. H.; SOULÉ, M. E. **Conservation and evolution**. New York: Cambridge University Press, 1981. 336 p.

GIBAS, C., JAMBECK, P. **Desenvolvendo bioinformática:** ferramentas de softwares para aplicação em biologia. Rio de Janeiro: Campus, 2001. 437 p.

GRANDBASTIEN, M. A. Retroelement in higher plants. **Trends in Genetics**, Cambridge, v. 8, n. 3, p. 103-108, Mar. 1992.

GRIFFITHS, A. J. F. et al. **Introdução a genética**. 9. ed. Rio de Janeiro: Guanabara Koogan, 2008. 764 p.

HAMILTON, J. P.; BUELL, C. R. Advances in plant genome sequencing. **Plant Journal**, Oxford, v. 70, n. 1, p. 177-190, Apr. 2012.

HARDON, J. J.; TAN, G. Y. Interespecific hibrids in the genus *Elaeis* I. crossability, cytogenetics and fertility of F1 hybrids of *E. guineensis* X *E. oleifera*. **Euphytica**, Wageningen, v. 18, n. 3, p. 372-379, Dec. 1969.

INTERNATIONAL ENERGY AGENCY. **World energy outlook 2010.** Paris: OECD, 2010.

JACCOUD, D. et al. Diversity arrays: a solid state technology for sequence information independent genotyping. **Nucleic Acids Research**, Oxford, v. 29, n. 4, p. 1-7, Jan. 2001.

KUMAR, A.; BENNETZEN, J. L. Plant retrotransposons. **Annual Review of Genomics and Human Genetics**, Palo Alto, v. 33, n. 1,p. 479-532, Dec. 1999.

KUSHAIRI, A.; RAJANAIDU, N. Breeding populations, seed production and nursey management. In: BARSISON, Y.; JALANI, B. S.; CHAN, K. W. **Advances in oil palm research**. Malaysian: Malaysian Palm Oil Board, 2000. p. 39-96.

LING, H. Q. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. **Nature**, London, v. 496, n. 3, p. 87-90, Mar. 2013.

LISCH, D. How importante are transposon for plant evolution? **Nature Reviews Genetics**, London, v. 14, n. 1, p. 49-61, Jan. 2011.

LITT, M.; LUTY, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. **The American Journal of Human Genetics**, Houston, v. 44, n. 3, p. 397-401, Mar. 1989.

LOW, E. T. et al. Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs: Identifying genes associated with callogenesis and embryogenesis. **BMC Plant Biology**, London, v. 8, n. 62, p. 1-19, May 2008.

MALOSETTI, M.; ABADIE, T. Sampling strategy to develop a core collection of Uruguayan maize landraces based on morphological traits. **Genetic Resources and Crop Evolution**, Dordrecht, v. 48, n. 4, p. 381-390, Aug. 2001.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual Review Genomics and Human Genetics**, Palo Alto, v. 9, n. 1, p. 397-402, Sept. 2008.

MARITA, J. M.; RODRIGUEZ, J. M.; NIENHUIS, J. Development of an algorithm identifying maximally diverse core collections. **Genetic Resources and Crop Evolution**, Dordrecht, v. 47, n. 5, p. 515-526, Oct. 2000.

MCCARTHY, E. M. et al. Long terminal repeat retrotransposons of *Oryza sativa*. **Genome Biology**, London, v. 3, n. 10, p. 1-11, Sept. 2002.

MEDINI, D. et al. Microbiology in the post-genomic era. **Nature Reviews Microbiology**, London, v. 6, n. 6, p. 419-430, June 2008.

MEYES, B. C.; TINGEY, S. V.; MORGANTE, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. **Genome Research**, New York, v. 11, n. 10, p. 1660-1676, Oct. 2001.

MINISTÉRIO DAS MINAS E ENERGIA. Boletim mensal combustíveis renováveis. Brasília: Ministério das Minas e Energia, 2012. 27 p. Disponível em: <http://www.mme. gov.br/spg/menu/publicacoes.html>. Acesso em: jan. 2015.

MORGANTE, M. et al. Gene duplication and exon shuffing by helitron-like transposons generate instraspecies diversity in maize. **Nature Genetics**, New York, v. 37, n. 9, p. 997-1002, Jul. 2005.

MULLIS, K. B., FALOONA, F. A. Specific synthesis of DNA in vitro a polymerase-catalyzed chain reaction. **Methods in Enzymology**, New York, v. 155, n. 4, p. 335-350, Nov. 1987.

MYLES, S. et al. Rapid genomic characterization of the genus vitis. **PLoS One**, San Francisco, v. 5, n. 1, p. 1-9, Jan. 2010.

OLIVEIRA, E. J. et al. Development of a cassava core collection based on single nucleotide polymorphism markers. **Genetics and Molecular Research**, Ribeirão Preto, v. 13, n. 3, p. 6472-6485, Aug. 2014.

RIOS, S. A. et al. **Recursos genéticos de palma de óleo (Elaeis guineensis Jacq.) e Caiaué (*Elaeis oleifera* (H. B. K.) Cortés).** Manaus: Embrapa Amazônia Ocidental, 2012. 39 p. (Documento, 96).

RONAGHI, M. Pyrosequencing shelds lights on DNA sequencing. **Genome Research**, New York, v. 11, n. 1, p. 3-11, Jan. 2001.

SANSALONI, C. P. **Desenvolvimento e aplicações de DArT (*Diversity Arrays Technology*) e genotipagem por sequenciamento (*Genotyping-by-Sequencing*) para análise genética em *Eucalyptus***. 2012. 145 p. Tese (Doutorado em Ciências Biológicas) – Universidade de Brasília, Brasília, 2012.

SCHWANT, R. L.; FOY, B. D.; PHOENIX, T. **Learning perl**. 6. ed. Sebastopol: O'Reilly Media, 2011. 388 p.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature Biotechnology**, New York, v. 26, n. 10, p. 1135-1145, Oct. 2008.

SHULAEV, V. et al. The genome of woodland strawberry. **Nature Genetics**, New York, v. 43, n. 2, p. 109-116, Feb. 2011.

SING, R. et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. **Nature**, London, v. 500, n. 7462, p. 335-339, July 2013a.

SING, R. et al. The oil palm SHELL gene controls oil yield and encodes a homologue od sedstick. **Nature**, London, v. 500, n. 7462, p. 340-344, July 2013b.

SPAGNOLETTIZEULI, P. L.; QUALSET, C. O. Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. **Theoretical and Applied Genetics**, Berlin, v. 87, n. 3, p. 295-304, Nov. 1993.

TAKAYAMA, K. et al. Relationships and genetic consequences of contrasting modes of speciation among endemic species of Robinsonia (Asteraceae, Senecioneae) of the Juan Fernández Archipelago, Chile, based on AFLPs and SSRs. **New Phytologist**, Lancaster, v. 205, n. 1, p. 415-428, Jan. 2015.

THE TOMATO GENOME CONSORTIUM. The tomato genome sequence provides insights into fleshy fruit evolution. **Nature**, London, v. 485, n. 7400, p. 635-641, May 2012.

TÓTH, G.; GÁSPÁRI, Z.; JURKA, J. Microsatellites in different eukaryotic genomes: survey and analysis. **Genome Research**, New York, v. 10, n. 7, p. 967-978, July 2000.

TRANBARGER, T. J. et al. Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism. **Plant Physiology**, Lancaster, v. 156, n. 2, p. 564-584, June 2011.

UNITED STATES DEPARTMENT OF AGRICULTURE. Disponível em: <http://www.usda.gov/wps/portal/usda/usdahome>. Acesso em: 26 jan. 2014.

URQUIAGA, S.; ALVES, B. J. R.; BOODEY, R. M. Produção de biocombustíveis a questão do balanço energético. **Revista de Política Agrícola**, Brasília, v. 14, n. 1, p. 42-46, Jan. 2005.

UTHAIPAISANWONQ, P. et al. Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). **Gene**, Amsterdam, v. 500, n. 2, p. 172-180, June 2012.

VICIENT, C. M. et al. Active retrotransposons are a common feature of grass genome. **Plant Physiology**, Lancaster, v. 125, n. 3, p. 1283-1292, Mar. 2001.

WILLIAMS, J. G. et al. DNA polymorphism amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, London, v. 18, n. 22, p. 6531-6535, Nov. 1990.

CAPÍTULO 2 Genômica comparativa e análise de repetições no *draft* do genoma de caiaué (*Elaeis oleifera*)

RESUMO

Este estudo teve como objetivos principais averiguar a representatividade do *draft* do genoma de *E. oleifera* produzido pela Embrapa contra o *draft* de *E. oleifera* e o genoma de *E. guineensis* públicos e identificar *in silico* as principais classes de repetições em *tandem* e de elementos transponíveis nesse *draft*. O *draft* analisado neste estudo foi desenvolvido com 130X coberturas na plataforma *Illumina HISeq* 2000, e foi comparado com o *draft* do genoma de *E. oleifera* e com o genoma de *E. guineensis* públicos por meio do *software* Nucmer. Foi feita uma análise de repetições em *tandem* com os *softwares Tandem Repeat Finder* e *Tandem Repeats Analysis Program*. Para identificação de elementos transponíveis foi realizada uma identificação *de novo* e classificação por similaridade com o RepBase e bancos públicos de repetições de plantas. Foi possível observar que 68% do *draft* foi alinhado com 79% do *draft* de *E. oleifera* público (com uma identidade média de 96%). Uma grande proporção do *draft* (73%) foi alinhada com alta identidade (94%) para 70% do genoma de *E. guineensis*. Foi possível identificar 328.879 locos de repetições em *tandem* e a principal classe identificada foi de dinucleotídeos (4,74%). Foram identificados 618.284 locos de elementos transponíveis em 84.398 *scaffolds*, o que representa 55% do *draft*. Dentre esses potenciais elementos transponíveis, os transposons classe I ou retrotranposons foram expressivamente mais abundantes que os transposons classe II, sendo observado respectivamente 62,71 e 6,86%. O *draft* de *E. oleifera* possui grande parte dos genomas com que foi comparado, representando grande parte de um genoma altamente complexo e utilizando tecnologia de sequenciamento de custo acessível. Mais da metade do *draft* é composto por repetições, principalmente retrotransposons, a identificação das regiões de repetições pode contribuir no auxílio para a montagem de uma nova versão desse *draft*.

Palavras-chave: Palma de óleo. Bioinformática. Repetições em *tandem*. Elementos transponíveis.

ABSTRACT

This study had as main objectives to evaluate the representativeness of the *draft* genome developed by Embrapa when compared to the public genomes of *E. oleifera* and *E. guineensis*, as well as to identify *in silico* the main classes of *tandem* repeats and transposable elements in this *draft*. The *E. oleifera* genome *draft* developed by Embrapa resulted from a 130X redundancy using the *Illumina* Platform, and was compared to the public genomes using the Nucmer *software*. An analysis of *tandem* repeats was performed using the *Tandem Repeat Finder* and *Tandem Repeats Analysis softwares*. To *de novo* identify the transposable elements and to classify them by similarity, the RepBase public banks of plant repeats were used. It was possible to see that 68% of the *draft* was aligned to 79% of the *E. oleifera* public genome (with an average identity of 96%). A total of 73% of the *draft* was aligned with high identity rate (94%) to 70% of the *E. guineensis* public genome. It was possible to identify 328.879 *tandem* repeat loci, being dinucleotides the main type present (4,74%). A total of 618.284 transposable elements loci were identified in 84.398 *scaffolds*, representing 55% of the *draft*. Among these potential transposable elements, transposons class I or retrotranposons were significantly more abundant than transposons class II, representing respectively 62.71 and 6.86% of the total. The draft of *E. oleifera* sampled much of the genomes that were compared, representing much of a highly complex genome with an affordable cost of sequencing technology. More than half of the draft consists of repetitions, especially retrotransposons, the identification of regions of repetitions can contribute in aid for the installation of a new version of this draft.

Keywords: Oil palm. Bioinformatics. Tandem repeats. Transposable elements

1 INTRODUÇÃO

O estudo da genômica da palma de óleo americana (*Elaeis oleifera*), também conhecida como caiaué, se dá principalmente para otimizar o processo de incorporação de características favoráveis no dendê (*E. guineensis*) via obtenção de híbridos interespecíficos (HIE). HIE entre essas duas espécies são resistentes/tolerantes ao amarelecimento fatal que constitui uma limitação no desenvolvimento da palmicultura no Brasil.

As subamostras de origem Manicoré (população natural no estado do Amazonas) é a população melhor representada no banco de germoplasma da caiaué da Embrapa Amazônia Ocidental, devido à qualidade das plantas e aos bons índices de germinação (RIOS et al., 2012). O DNA de um indivíduo dessa região foi utilizado no sequenciamento pela plataforma *Illumina HISeq* 2000 (130X de cobertura) para desenvolver o *draft* do genoma de caiaué que foi analisado neste estudo.

A comparação de genomas pode ser utilizada de diferentes formas, desde avaliar relações evolutivas entre espécies até a de busca de divergência entre indivíduos de uma mesma espécie. Essa análise também pode ser usada para avaliar a qualidade de um genoma rascunho em relação a um genoma com uma montagem mais representativa da espécie. O *draft* de *E. oleifera* e o genoma de *E. guineensis* (SING et al., 2013) foram montados com uma elevada cobertura (25 e 26X respectivamente) de *reads* gerados pela tecnologia de sequenciamento Roche/454 que promove leituras maiores (em torno de 700 pares de base) e consequentemente permitem uma montagem mais precisa, desta forma esses dois genomas podem ser usados como base de avaliação de representatividade do *draft* avaliado.

Os genomas eucariotos são densamente constituídos por regiões repetitivas, principalmente repetições em *tandem* e elementos de transposição

(TE's). São exemplos do primeiro grupo os microssatélites, minissatélites e telômeros; do segundo grupo, os elementos genéticos móveis ou simplesmente, transposons. Os microssatélites são utilizados como marcadores moleculares em estudos de melhoramento genético de plantas e testes de paternidade (SHARMA; GROVER; KAHL, 2007); já os transposons têm aplicações na regulação gênica e terapia genética (IVICS; IZSVAK, 2006).

De acordo com os tipos de transposição, mediada por RNA ou DNA, os TE's podem ser divididos em duas classes principais. Elementos de classe I, ou retrotransposons, é o tipo mais comum de elemento genético móvel em genomas eucariotos e se transpõem através de um intermediário de RNA que é transcrito novamente em DNA e se integra em outra posição no genoma (FESCHOTTE; JIANQ; WESSLER, 2002). Elementos de classe II, ou transposons de DNA, utilizam um mecanismo de corte do elemento em determinada posição do genoma e integração deste em outra região (BOWEN; JORDAN, 2002).

A caracterização e mascaramento dessas repetições são de fundamental importância na compreensão do genoma analisado e no processo de predição gênica respectivamente.

Este estudo teve como objetivos de: (i) averiguar a representatividade do *draft* do genoma de *E. oleifera* produzido pela Embrapa com o *draft* de *E. oleifera* e o genoma de *E. guineensis* públicos e (ii) identificar *in silico* as principais classes de repetições em *tandem* e de elementos transponíveis nesse *draft*.

2 MATERIAL E MÉTODOS

2.1 *Draft* do genoma de caiaué

A montagem do *draft* foi realizada nos laboratórios de bioinformática da Embrapa Agroenergia e da Embrapa Informática Agropecuária, utilizando a ferramenta ALLPATHS-LG (GNERRE et al., 2011). Foram realizados controles de qualidade por meio dos programas FASTX-Toolkit e FASTQC. *Scripts* em Perl foram usados para a manipulação e formatação dos dados gerados. A versão inicial dessa montagem produziu um total de 85.612 de *scaffolds* com um N50 de 27K (*scaffolds* com *gaps*) de uma cobertura de 130X de *reads* produzidos pelo sequenciador *Illumina HiSeq* 2000 (Tabela 1).

2.2 Comparação genômica

Foi utilizado no alinhamento o *draft* de *E. oleifera* e o genoma de *E. guineensis* públicos (SING et al., 2013). O *software* utilizado para o alinhamento foi o Nucmer (*-maxmatch*) que faz parte do pacote de *softwares* MUMmer 3.23 (KURTZ et al., 2004). Os utilitários *delta-filter* (*-q*), *show-coords* (*-rcl*) e *dnadiff* (parâmetros padrão) foram usados na filtragem dos alinhamentos, obtenção das coordenadas de mapeamento e na geração do relatório estatístico do alinhamento respectivamente.

Foi desenvolvido um *script* em Perl (Apêndice A) para agregar alinhamentos menores que estavam inseridos nas coordenadas de outro alinhamento, assim permanecendo somente regiões únicas de alinhamento de um *scaffold*.

Tabela 1 Sumário da montagem do *draft* do genoma de *Elaeis oleifera*

Montagem	**Resultado**
Número total de *reads*	1.319.238.018
GC nos *reads* (%)	37,2
Tamanho mínimo de *contig* (pb)	1.000
Número de *contigs*	163.136
Número de *scaffolds*	85.612
Tamanho total de *contigs* (pb)	947.205.433
Tamanho total de *scaffolds* com *gaps* (pb)	1.015.396.310
Tamanho de *contigs* N50 (kb)	10,6
Tamanho de *scaffolds* N50 (kb)	25
Tamanho de *scaffolds* N50 (kb) com gaps	27
Número de *scaffolds* por Mb	84,31
Tamanho de *gaps* em *scaffolds*	750
Estimativa do tamanho do genoma (pb)	1.503.187.609

2.3 Análise de repetições no *draft*

Para a análise de repetição em *tandem* foram seguidas as seguintes etapas: (1) identificação, (2) análise e (3) classificação das repetições. Já as análises de elementos transponíveis foram realizadas baseadas na identificação de novo e classificação por similaridade.

2.3.1 Identificação de repetições em *tandem*

Para a identificação de repetições em *tandem* no *draft* de *E. oleifera* foi utilizado o *software Tandem Repeat Finder* – TRF (BENSON et al., 1999) com os seguintes parâmetros *match* 2, *mismatch* 7, *delta* 7, *PM 80*, *PI 10*, *minscore* 50, *maxperiod500*, *–f*, *–d* e *–m*. Para a análise e classificação das repetições identificadas, foi utilizado o *software Tandem Repeats Analysis Program* – TRAP (SOBREIRA; DURHAM; GRUBER, 2006) com os parâmetros padrão.

2.3.2 Identificação de Elementos Transponíveis (TE's)

Primeiro foi utilizado o *software* Repeat Modeler (parâmetros padrão) que compõe um pipeline com os *softwares* RECON (BAO; EDDY, 2002), Repeat Scout (PRICE; JONES; PEVZNER, 2005), Repeat Masker, TRF (BENSON et al., 1999) e RMBLAST para a identificação de elementos transponíveis (TE's) *de novo* no *draft* de *E. oleifera*. Retrotransposons do tipo LTR (*Long terminal repeats*) foram identificados com o *software* LTR_FINDER (XU; WANG, 2007) com os parâmetros padrão. Todas as repetições com tamanho >100 pb constituíram a biblioteca de TE's *de novo* de *E. oleifera*.

2.3.3 Classificação da Biblioteca de TE's de novo

A classificação da biblioteca de TE's *de novo* foi feita através de BLASTN (e-value ≤ 1e-5, identidade ≥ 70% e tamanho mínimo de alinhamento ≥ 80 pb) contra o RepBase e contra o banco público de repetições de plantas MIPS *Repeat database* que integra outros bancos públicos (TREP, TIRG *repeats*, *PlantSat* e *Genbank*). TE's identificados que não puderam ser classificados foram atribuídos como retrotransposons não classificados ou DNA transposons não classificados.

2.3.4 Anotação de TE's

Foi utilizado para a busca das coordenadas de TE's no *draft* de *E. oleifera* o *software Repeat Masker* com uma biblioteca customizada (combinação de repetições do Rep Base, MIPs e a biblioteca de TE's *de novo*). Esse *software* também foi utilizado para gerar uma versão do *draft* com as regiões de repetição mascaradas.

3 RESULTADOS

3.1 Comparação genômica

O resultado da comparação genômica mostra grande semelhança entre as sequências avaliadas. Foram alinhados 68% do *draft* avaliado, o que corresponde 680 megabase (Mb) do total de bases, contra 78% (365 Mb) do *draft* de *E. oleifera* público, com 96% de identidade (Figura 1).

Um total de 730 Mb do *draft* de caiaué (73%) foi alinhado, com alta identidade (94%), contra 70% (460 Mb) do genoma de *E. guineensis* (Figura 2). Para o genoma de *E. guineensis*, foi analisado a cobertura do *draft* para os 16 cromossomos, que variou de 69 a 77% (Figura 2).

3.2 Identificação de repetições em *tandem*

Foi possível identificar 328.879 locos de repetições em *tandem* no *draft* de caiaué, o que corresponde a um total de 35 Mbe que representa 0,35% do total de bases desse *draft*. Foram excluídos 23.056.935 pares de base (pb) devido à redundância das repetições. O número do motivo das repetições variou de 1 a 501.

As principais classes de sequências repetitivas identificadas foram mononucleotídeos (0,92%), dinucleotídeos (4,74%), trinucleotídeos (1,05%), tetranucleotídeos (1,13%), pentanucleotídeos (1,73%), hexanucleotídeos (1,37%) e outras (89,06%) (Figura 3). Para cada classe, os principais motivos de repetição foram $(T/A)_n$ – 99,97%, $(AT)_n$ – 48,05%, $(TTA)_n$ – 42,22%, $(ACAT)_n$ – 47,53%, $(TATAT)_n$ – 30,86% e $(TTTTTC)_n$ – 32,68% respectivamente. Entre as principais classes analisadas, as repetições mais abundantes foram os dinucleotídeos, com 15.574 locos identificados (Tabela 2).

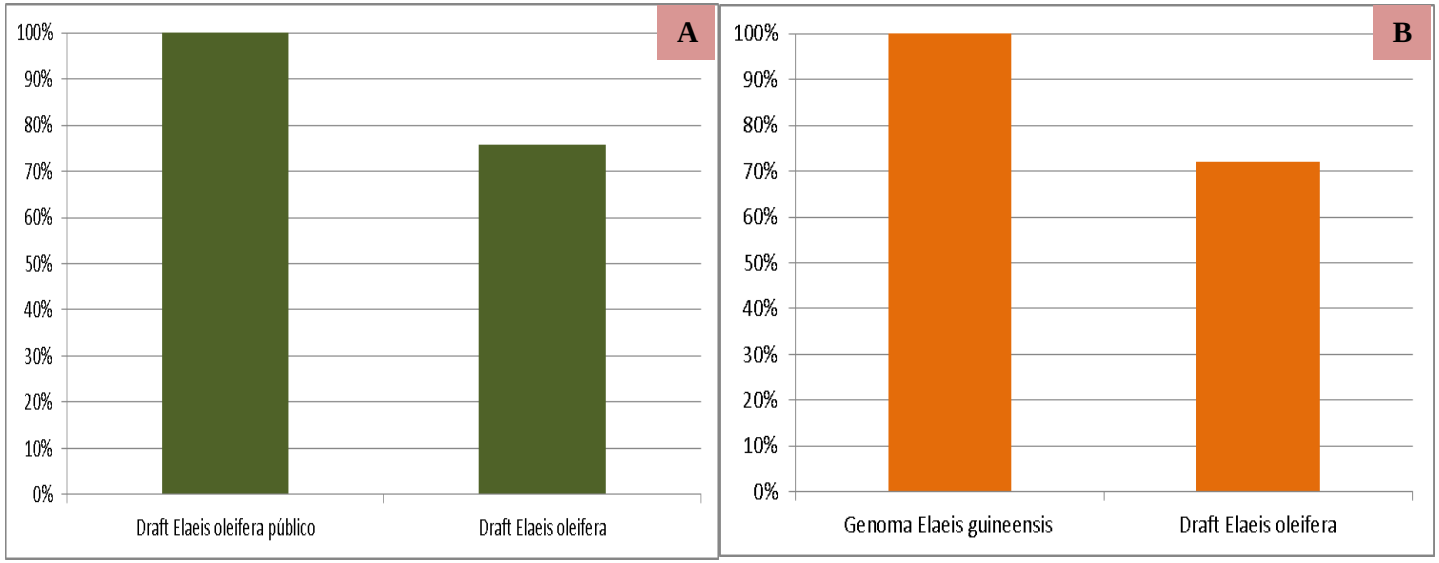

Figura 1 Comparação genômica entre o *draft* de caiaué (*Elaeis oleifera*) desenvolvido pela Embrapa, com 130X coberturas do sequenciador *Illumina Hiseq* 2000, contra o *draft* de *E. oleifera* público (A) e contra o genoma de *E. guineensis* público (B)

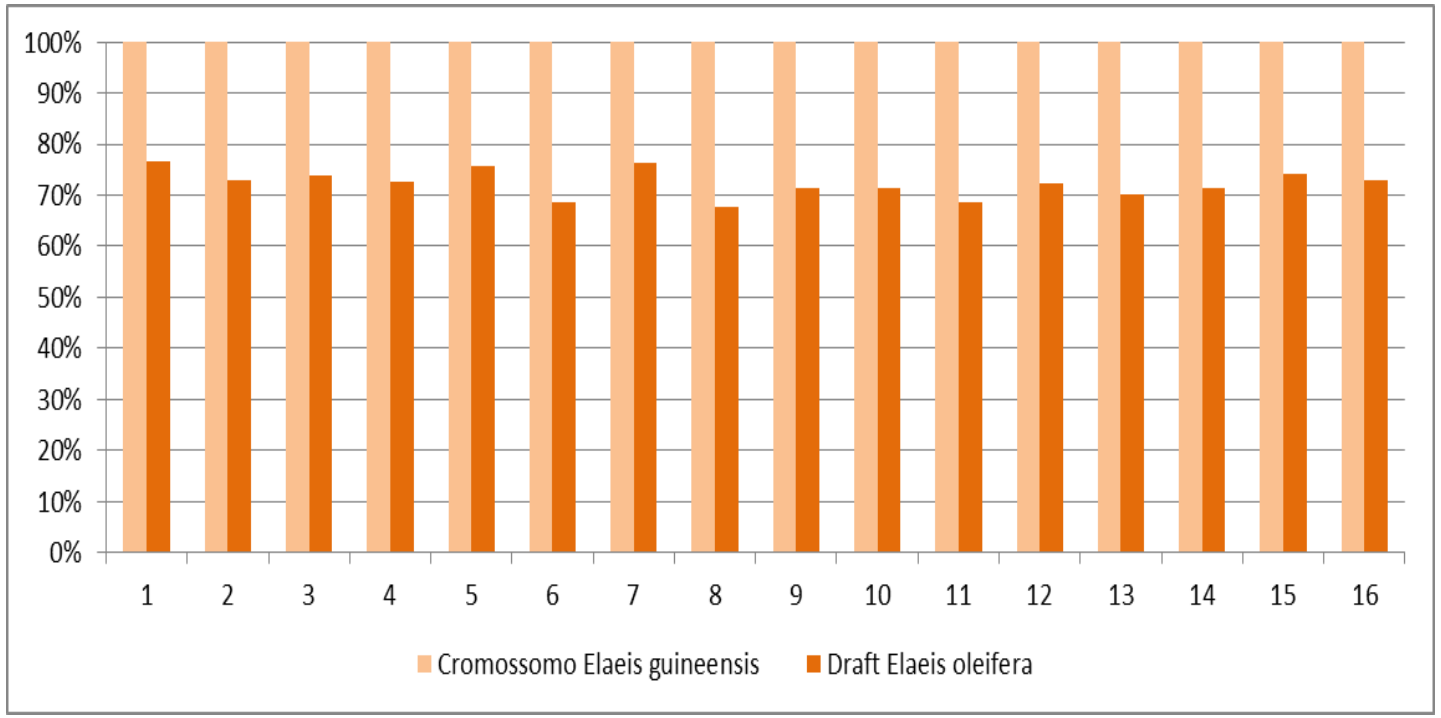

Figura 2 Comparação da cobertura do *draft* de caiaué (*Elaeis oleifera*), desenvolvido pela Embrapa, em relação aos 16 cromossomos do genoma De *Elaeis guineenses* público

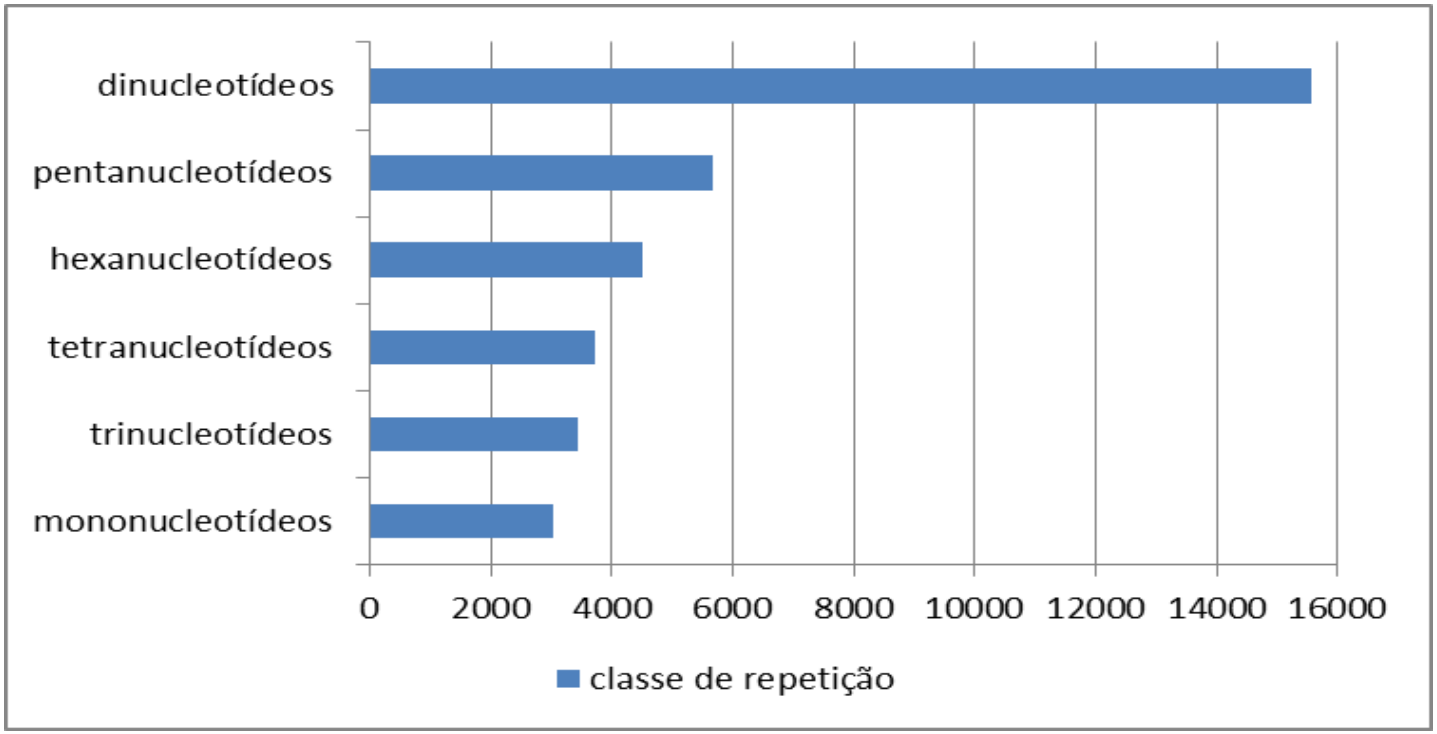

Figura 3 Distribuição das principais classes de repetição em *tandem* identificadas no *draft* do genoma de caiaué (*Elaeis oleifera*) desenvolvido pela Embrapa

Tabela 2 Motivo da repetição das principais classes de repetição em *tandem* identificadas no *draft* do genoma de *Elaeis oleifera*

Classe de Repetição	N. de Sequências
Mononucleotídeos	
(T/A)n	3.017
(C/G)n	4
Dinucleotídeos	
(AT)n	7.483
(CT)n	6.306
Trinucleotídeos	
(TTA)n	1.456
(AAG)n	1.304
Tetanucleotídeos	
(AAAG)n	1.774
(TATAT)n	818
Pentanucleotídeos	
(TATAT)n	1.754
(TTTCT)n	1.719
Hexanucleotídeos	
(TTTTTC)n	1.473
(ATTAAT)n	439

3.3 Identificação de Elementos transponíveis

Foram identificados 618.284 locos de TE's em 84.398 dos *scaffolds* do *draft*, o que totaliza 550 Mb e representa 55% do total de bases desse *draft* (Figura 4). Dentre esses potenciais elementos transponíveis, os transposons classe I ou retrotranposons foram expressivamente mais abundantes que os transposons classe II, sendo observado respectivamente 62,71 e 6,86%.

Dentre os elementos transponíveis classe I identificados, a principal ordem foi de LTR (*Long Terminal Repeats*). As duas principais superfamílias foram Copia (21%) e Gypsy (5%), e 4% foram de outras superfamílias. Não apresentaram similaridades com repetições conhecidas e não foram classificadas 29% dos LTRs. Embora em menor proporção

também, foi possível identificar outros tipos de retroelementos, tais como, LINE – *Long Interspeted Nuclear ElementS* (<1%) e SINES – *Short Interspeted Nuclear ElementS* (1%).

Para os TE's classe II identificados, a principal superfamília foi Tir-hAT (1%), houve uma pouca proporção de CACTA (menor que 1%) e não foram identificados MITES. Foram identificadas 2% de repetições específicas de *E. guineensis* (DRepEG). Uma grande proporção de TE's identificados no *draft* não é conhecida (30%). A Tabela 3 apresenta um comparativo dos TEs identificados no *draft* de *E. oleifera* em relação ao genoma de *E. guineensis*. O *draft* de *E. oleifera* teve todas as regiões identificadas de TE's mascaradas pelo *software* Repeat Masker.

Tabela 3 Comparação da porcentagem de elementos transponíveis identificados no *draft* do genoma de *Elaeis oleifera* e no genoma de *E. guineensis*

Elementos Transponíveis	***Elaeis oleifera* (%)**	***Elaeis guineensis* (%)**
LTR-Copia	21	33
LTR-Gypsy	5	8%
Outros LTR	4	6%
LINE	<1	<1
SINE	1	-
DRepEG	2	7
DNA Transposon	7	4
Não classificado	59	40
Total (Mb)	550	282.3

LTR - *Long Terminal Repeats*; LINE – *Long Interspeted Nuclear Element*; SINE – *Short Interspeted Nuclear Element;* DRepEG – repetições não caracterizadas de *Elaeis guineensis*; Mb – megabase.

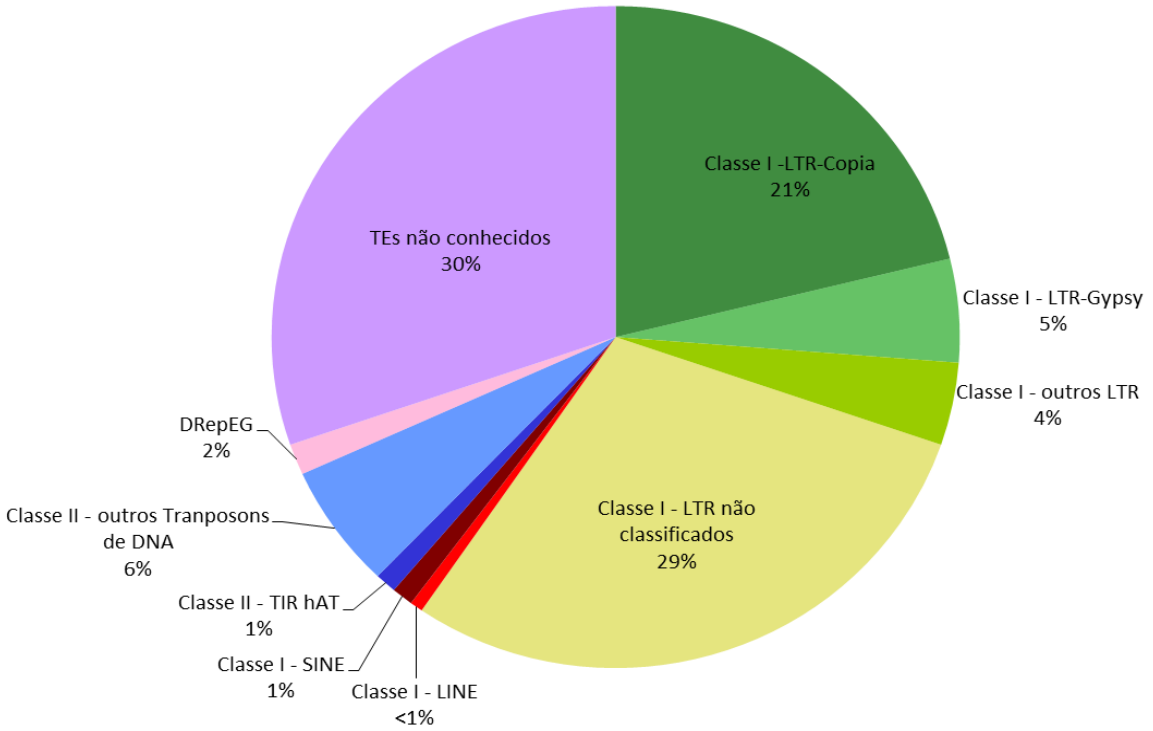

Figura 4 Principais classes identificadas de elementos transponíveis no *draft* do genoma de *Elaeis oleifera*

Legenda: LTR - *Long Terminal Repeats*; LINE – *Long* ***interspersed*** *Nuclear Element*; SINE – *Short* ***interspersed*** *Nuclear Element;* DRepEG – repetições não caracterizadas de *E. guineensis.*

4 DISCUSSÃO

Esse estudo apresenta uma análise de comparação genômica entre genomas de *Elaeis* spp., como também análise de elementos repetitivos em uma versão inédita de um *draft* do genoma de caiaué (*E. oleifera*), desenvolvido com base na plataforma *Ilumina HISeq* 2000 com um genótipo oriundo da população Manicoré da Amazônia. A população Manicoré é de grande interesse no programa de melhoramento genético da palma de óleo desenvolvido pela Embrapa, e que busca o desenvolvimento de híbridos interespecíficos com E. guineensis.

4.1 Comparação genômica

Os resultados encontrados neste trabalho indicam que o *draft* de *Elaeis oleifera* da Embrapa apresenta grande parte da montagem dos genomas públicos. O que corrobora que a tecnologia de sequenciamento *Illumina Hiseq* 2000 foi eficiente na amostragem de um genoma altamente complexo de uma planta alógama, que apresenta alto grau de heterozigose. A maior vantagem na utilização dessa estratégia foi conseguir representar uma grande parte do genoma dessa espécie por um custo acessível. Outros genomas vegetais já foram sequenciados por meio da tecnologia de sequenciamento *Illumina* (LING et al., 2013; SHULAEV et al., 2011).

A plataforma *Illumina Hiseq* possui alta acurácia (98%) com a geração de um grande volume de dados (600 Gb/corrida) e possui um baixo custo ($0,07/1 Mb), no entanto, sua maior desvantagem é a montagem com *reads* curtos (100 pb) (LIU et al., 2012). A opção análoga para o sequenciamento seria a tecnologia Roche/454 que gera leituras maiores, em torno de 700 pb, mas

possui alto custo ($10/1 Mb) e uma elevada taxa de erro no sequenciamento de homopolímeros (LIU et al., 2012).

O genoma de *E. guineensis* e *E. oleifera* foi estimado por citometria de fluxo em um tamanho de 4,32 ± 0,173 e 4.43 ± 0.018 pg (picograma), respectivamente (CAMILLO et al., 2014), o que corresponde a aproximadamente 2 gigabase - Gb para ambos os genomas. O *draft* de *E. oleifera* analisado nesse trabalho possui uma montagem final de aproximadamente 1 Gb e conseguiu amostrar mais de 70% do *draft* público. No entanto, devido à dificuldade de montar sequências repetitivas adjacentes com *reads* curtos, muitas sequências podem estar montadas de forma errada e esse tamanho final pode estar superestimado.

Grande parte do *draft* de *E. oleifera* avaliado (72,83%) foi alinhada contra o genoma de *E. guineensis*, evidenciando a alta relação entre essas espécies que possuem compatibilidade reprodutiva. Uma menor proporção (68,24%) foi alinhada contra o *draft* de *E. oleifera* público. Esse resultado indica uma montagem mais representativa do genoma de *E. guineenis*, e que esse portanto pode ser utilizado com maior confiabilidade, mesmo em estudos de genômica com *E. oleifera*.

Um genoma de uma espécie na verdade é a representação de um organismo individual ou linhagem padrão da qual o DNA foi obtido. Essa sequência servirá então como sequência referência para a espécie. Assim, nenhuma sequência de genoma representa verdadeiramente o genoma de toda a espécie. Para o sequenciamento do *draft* de caiaué analisado nesse trabalho foi utilizado o DNA de uma planta da população Manicoré da Amazônia. Essa população é de interesse no melhoramento da palma de óleo da Embrapa, sendo utilizada como parental feminino para gerar a cultivar BRS Manicoré tolerante/resistente ao AF (CUNHA; LOPES, 2010). Devido a esse fato, uma nova versão desse *draft* será produzida agregando mais informação de

sequenciamento, e com o conhecimento das regiões de repetição encontradas nesse estudo será possível realizar uma montagem mais precisa dessa nova versão.

A caracterização completa e refinamento desse *draft* serão úteis em diversos aspectos no melhoramento genético de, principalmente no desenvolvimento e caracterização de marcadores moleculares, estudos filogenéticos, sintenia gênica, mapeamento físico, predição gênica, seleção assistida por marcadores, seleção genômica ampla, entre outras aplicações. Um dos principais desafios do melhoramento genômico é a integração de metodologias e técnicas empregadas no melhoramento genético convencional com as tecnologias e estratégias de biologia molecular. Isso inclui, por exemplo, o emprego de informação genética derivada de mapas de ligação construídos com marcadores moleculares.

4.2 Repetições em *tandem*

Foram identificados nesse estudo 328.879 locos de repetições em *tandem*, que são potenciais marcadores microssatélites para *E. oleifera*. Marcadores microssatélites destacam-se por ser multialélicos, codominantes e de alta reprodutibilidade (TÓTH; GÁSPÁRI; JURKA, 2000). Poucos trabalhos de desenvolvimento e aplicação de marcadores microssatélites estão disponíveis para o gênero *Elaeis* spp. (BILLOTTE et al., 2010; TING et al., 2010).

A grande limitação dessa classe de marcadores reside na necessidade de isolamento e desenvolvimento de *primers* específicos para cada espécie, sendo esse processo demorado, trabalhoso e de alto custo. Entretanto, essa tarefa vem sendo facilitada com o sequenciamento de DNA de diferentes espécies, que resultou na disponibilidade de milhares de sequências em bancos públicos e privados de DNA. As milhares de sequências, aliadas às novas técnicas de

bioinformática, permitiram que o desenvolvimento de *primers* microssatélite ocorresse sem a necessidade de todas as etapas iniciais da estratégia tradicional (OSTRANDER et al., 1992), tornando o processo menos trabalhoso.

A maior proporção de classes de repetição em *tandem* identificadas no *draft* analisado foi o de dinucleotídeos. A maioria dos microssatélites (48-67%) estudados são dinucleotídeos (WANG et al., 1994). Dentro da classe de dinucleotídeos o motivo mais frequente identificado foi AT, seguido por CT. Essas observações estão de acordo com estudos feitos em maçã (HAN; KORBAN, 2008), *Arabidopsis* (TAMANNA; KHAN, 2005), soja (SHULTZ et al., 2007), mamão (LAI et al., 2006), dentre outras plantas e mostram que motivos ricos em AT são muito mais prevalentes nos genomas das plantas superiores.

4.3 Elementos transponíveis

Mais de 50% do *draft* de *E. oleifera* é composto por elementos transponíveis. Esse fato correlaciona com o paradoxo do valor C em que o tamanho do genoma em organismos eucariotos está relacionado com a quantidade de regiões repetitivas e não com o conteúdo gênico. Genomas pequenos como o de *Arabidopsis thaliana* possuem somente 10% de DNA repetitivo, este valor é muito maior em outros genomas de plantas sequenciados, como álamo – 42% (TUSKAN et al., 2006), mamão – 51,9 % (MING et al., 2012), maçã – 42,4% (VELASCO et al., 2010) e dendê – 43% (SING et al., 2013).

Alguns domínios dos cromossomos têm uma cromatina altamente condensada, chamada de heterocromatina. Outros domínios são embalados em uma cromatina menos condensada, chamada de eucromatina. Já há algum tempo, os geneticistas sabem que o DNA da heterocromatina continha poucos

genes, enquanto a eucromatina é rica em genes. A heterocromatina é composta essencialmente por regiões repetitivas, às vezes chamadas de DNA lixo (OHNO, 1972). Assim os cromossomos densamente compactados de heterocromatina foram ditos formando uma estrutura que era inacessível a proteínas reguladoras e não apropriada para atividade gênica. Por esse motivo, em um trabalho de anotação genômica, é essencial para a predição gênica realizar identificação, anotação e mascaramento de regiões repetitivas.

Dos TE's identificados neste estudo, 59% são do tipo classe I de retrotransposons. Esse resultado já era esperado já que as grandes diferenças no tamanho dos genomas das espécies de plantas estão, geralmente, associadas à presença de diferentes quantidades de retrotransposons. Quanto maior o genoma vegetal, maior é a chance de este conter uma grande quantidade de retroelementos. Por exemplo, genomas grandes, como o da cevada, são compostos por até 70% desses elementos (VICIENT et al., 2001), enquanto em genomas pequenos, como o do arroz, esses representam apenas 17% da composição do genoma (MCCARTHY et al., 2002).

Dentro da classe I, houve uma maior presença da ordem LTR (*Long Terminal Repeat*) e as duas superfamílias que mais se destacaram foram a do tipo Copia (21%) e Gypsy (5%); o que parece ser típico em genomas de monocotiledôneas (DU et al., 2010). Houve baixa proporção de LINES (<1%) e SINES (1%) já que esse tipo de elemento parece ser mais abundante em genomas animais do que em vegetais (WICKER et al., 2007).

Os elementos de classe II estão pouco presente no *draft* do genoma de *E. oleifera* (7%), e a superfamília com o maior destaque foi a hAT (1%). Os membros da superfamília hAT são encontrados tantos em monocotiledôneas, como os da conhecida família Ac-Ds em milho (FEDOROFF; WESSLER; SHURE, 1983), quanto em dicotiledôneas, como os da família Tag1 em *Arabidopsis thaliana* (TSAY et al., 1993).

Um fato interessante foi a alta proporção de elementos não classificados (59%), e de 2% de repetições identificadas, mas não caracterizadas no genoma de *E. guineensis* – DrepEG (CASTILHO; VERSHININ; HESPLOP-HARRISON, 2000). Esse fato pode ser explicado devido aos bancos de repetições de espécies monocotiledôneas próximas ainda não estarem bem descrito.

A eficiência do método utilizado na identificação e anotação dos TE's foi comprovada devido à semelhança dos valores em porcentagem dos TE's identificados no *draft* de *E. oleifera* e daqueles encontrados para os do genoma de *E. guineensis* (Tabela 3). No entanto, embora o método geral tenha seguido os mesmos princípios, as ferramentas para a construção da biblioteca *de novo* desse genoma (SING et al., 2013) foram diferentes das usadas neste estudo.

Elementos transponíveis ativos exercem grande influência na expressão gênica e na evolução do genoma em plantas (LISCH, 2013). Sabe-se que os organismos desenvolveram mecanismos de equilíbrio dos efeitos da transposição, tais como a transposição para genes silenciados ou metilação do DNA (abrigos seguros) e remodelamento de histonas pelo organismo. A diversidade genética criada por elementos genéticos móveis é uma importante fonte de variação funcional sobre a qual a seleção atua ao longo do tempo evolutivo. Esses elementos são de grande importância na compreensão da função do DNA não expresso, ou lixo, até pouco tempo atrás considerado como de pouca importância pelos cientistas.

5 CONCLUSÕES

A montagem do *draft* do genoma de caiaué realizada com base na plataforma *Illumina* promoveu uma alta amostragem (mais de 70%) em relação aos genomas avaliados, representando grande parte de um genoma altamente complexo com uma tecnologia de sequenciamento de custo acessível.

Uma grande porção (55%) do *draft* estudado é composta por repetições tipicamente identificadas em genomas vegetais. Devido à complicação da montagem de repetições, o conhecimento dessas regiões poderá contribuir no aprimoramento da montagem de uma nova versão desse *draft*.

REFERÊNCIAS

BAO, Z.; EDDY, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. **Genome Research**, New York, v. 12, n. 8, p. 1269-1276, Aug. 2002.

BENSON, G. et al. Tandem repeats finder: a program to analyze DNA sequence. **Nucleic Acids Research**, London, v. 27, n. 2, p. 573-580, Jan. 1999.

BILLOTE, N. et al. QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). **Theoretical and Applied Genetics**, Berlin, v. 120, n. 8, p. 1673-1687, May 2010.

BOWEN, N. J.; JORDAN, I. K. Transposable elements and the evolution of eukaryotic complexity. **Current Issues Molecular Biolology**, Wymondham, v. 4, n. 3, p. 65-76, July 2002.

CAMILLO, J. et al. Reassessment of the genome size in *Elaeis guineensis* and *Elaeis oleifera*, and its interspecific hybrid. **Genomics Insights**, Boston, v. 7, n. 1, p. 13-22, May 2014.

CASTILHO, A.; VERSHININ, A.; HESPLOP-HARRISON, J. S. Repetitive DNA and the chromosomes in the genome of oil palm (*Elaeis guineensis*). **Annals of Botany**, London, v. 85, n. 6, p. 837-844, Feb. 2000.

CUNHA, R. N. V.; LOPES, R. **BRS Manicoré:** híbrido interespecífico entre o caiaué e o africano recomendado para áreas de incidências de amarelecimento-fatal. Manaus: Embrapa Amazônia Ocidental, 2010. 3 p.

DU, J. et al. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. **The Plant Journal**, Indiana, v. 63, n. 4, p. 584-598, May 2010.

FEDOROFF, N.; WESSLER, S.; SHURE, M. Isolation of the transposable maize controlling elements Ac and Ds. **Cell**, Cambridge, v. 35, n. 1, p. 235-242, Nov. 1983.

FESCHOTTE, C.; JIANQ, N.; WESSLER, S. R. Plant transposable elements: where genetics meet genomics. **Nature Review Genetics**, London, v. 3, n. 5, p. 329-341, May 2002.

GNERRE, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. **Proceedings of the National Academy of Sciences**, Washington, v. 108, n. 4, p. 1513-1521, Jan. 2011.

HAN, Y.; KORBAN, S. S. Na overview of the apple genome through BAC end sequence analysis. **Plant Molecular Biology**, Dordrecht, v. 67, n. 8, p. 581-589, Aug. 2008.

IVICS, Z.; IZSVÁK, Z. Transposons for gene therapy! **Current Gene Therapy**, France, v. 6, n. 5, p. 593-607, Oct. 2006.

KURTZ, S. et al. Versatile and open software for comparing large genomes. **Genome Biology**, London, v. 5, n. 2, p. 1-12, Jan. 2004.

LAI, C. W. et al. Analysis of papaya BAC end sequence reveals first insights into the organization of a fruit tree genome. **Molecular Genetics and Genomics**, Berlin, v. 276, n. 1, p. 1-12, July 2006.

LIU, L. et al. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, Cairo, v. 2012, n. 1, p. 1-11, Apr. 2012.

LING, H. Q. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. **Nature**, London, v. 496, n. 3, p. 87-90, Mar. 2013.

LISCH, D. How important are tranposons for plant evolution? **Nature Review Genetics**, London, v. 14, n. 1, p. 49-61, Jan. 2013.

MCCARTHY, E. M. et al. Long terminal repeat retrotransposons of *Oryza sativa*. **Genome Biology**, London, v. 3 , n. 10, p. 1-11, Sept. 2002.

MING, R. et al. Genome of papaya, a fast growing tropical fruit tree. **Tree Genetics and Genomics**, New York, v. 8, n. 3, p. 445-462, Jun. 2012.

OHNO, S. So much "junk" DNA in our genome. **Brookhaven Symposia in Biology**, Upton, v. 23, n. 2, p. 366-370, July 1972.

OSTRANDER, E. A. et al. Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequence. **Proceedings of the National Academy of Sciences**, Washington, v. 89, n. 8, p. 3419-3423, Apr. 1992.

PRICE, A. L.; JONES, N. C.; PEVZNER, P. A. De novo identification of repeat families in large genomes. **Bioinformatics**, Oxford, v. 21, n. 1, p. 351-359, June 2005.

RIOS, S. A. et al. **Recursos genéticos de palma de óleo (Elaeis guineensis Jacq.) e Caiaué (*Elaeis oleifera* (H. B. K.) Cortés).** Manaus: Embrapa Amazônia Ocidental, 2012. 39 p. (Documento, 96).

SHARMA, P. C.; GROVER, A.; KAHL, G. Mining microsatellites in eukaryotic genomes. **Trends in Biotechnology**, Amsterdam, v. 25, n. 11, p. 490-498, Oct. 2007.

SHULAEV, V. et al. The genome of woodland strawberry. **Nature Genetics**, New York, v. 43, n. 2, p. 109-116, Feb. 2011.

SHULTZ, J. L. et al. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic map of soybean. **Theoretical and Applied Genetics**, Berlin, v. 114, n. 6, p. 1081-1091, Apr. 2007.

SING, R. et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. **Nature**, London, v. 500, n. 7462, p. 335-339, July 2013.

SOBREIRA, T. J.; DURHAM, A. M.; GRUBER, A. TRAP: automated classification, quantification and annotation of tandemly repeated sequences. **Bioinformatics**, Oxford, v. 22, n. 3, p. 361-363, Feb. 2006.

TAMANNA, A.; KHAN, A. U. Mapping and analysis of simple sequence repeats in the *Arabidopsis thaliana* genome. **Bioinformatics**, Oxford, v. 1, n. 2, p. 64-72, Nov. 2005.

TING, N. C. et al. SSR mining in oil palm EST database: application in oil palm germplasm diversity studies. **Journal of Genetics**, Bangalore, v. 89, n. 2, p. 135-143, Aug. 2010.

TÓTH, G.; GÁSPÁRI, Z.; JURKA, J. Microsatellites in different eukaryotic genomes: survey and analysis. **Genome Research**, New York, v. 10, n. 7, p. 967-978, July 2000.

TSAY, Y. F. et al. Identification of a mobile endogenous transposon in *Arabinopsis thaliana*. **Science,** Washington, v. 260, n. 5106, p. 342-344, Apr. 1993.

TUSKAN, G. A. et al. The genome of black cottonwood, populus trichocarpa (Torr. & Gray). **Science,** Washington, v. 313, n. 5793, p. 1596-1603, Sept. 2006.

VELASCO, R. et al. The genome of the domesticated apple (Malus X domestica Borkh). **Nature Genetics**, New York, v. 42, n. 10, p. 833-839, Nov. 2010.

VICIENT, C. M. et al. Active retrotransposons are a common feature of grass genomes. **Plant Physiology**, Lancaster, v. 125, n. 3, p. 1283-1292, Mar. 2001.

WANG, Z. et al. Survey of plant short tandem DNA repeats. **Theoretical and Applied Genetics**, Berlin, v. 88, n. 1, p. 1-6, Apr. 1994.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, London, v. 8, n. 12, p. 973-982, Dec. 2007.

XU, Z.; WANG, H.; LTR_FINDER: an eficient tool the predictin of full-length LTR retrotransposons. **Nucleic Acids Research**, London, v. 35, n. 7, p. 265-273, May 2007.

CAPÍTULO 3 Caracterização genômica de marcadores por meio de genotipagem por sequenciamento (GBS) via plataforma DArTSeq e estabelecimento de coleção nuclear em caiaué (*Elaeis oleifera*)

RESUMO

Os principais objetivos desse estudo foram identificar e caracterizar marcadores PAVs (*presence/absence variants*) e SNPs (*single nucleotide polymorphism*) de caiaué (*Elaeis oleifera*) com base no mapeamento de sequências polimórficas ao genoma de dendê (*E. guineensis*) e delinear coleções nucleares para *E. oleifera*. Um banco de sequências gerados pela plataforma DArTSeq para 553 indivíduos de *E. oleifera* foi mapeado contra o genoma público de dendê com o *software* BWA, sendo que o pacote de *software* SAMtools foi utilizado para identificar os SNPs. Foram mapeados no genoma os modelos gênicos de tamareira (*Phoenix dactylifera*). O genoma foi dividido em intervalos de 5Mb para uma análise da distribuição dos marcadores e modelos gênicos. A partir de 1.666 SNPs, foram selecionados 500 com o maior índice de diversidade de Shannon para o estabelecimento de coleções nucleares a partir de 553 indivíduos representando 206 subamostras de *E. oleifera*, baseado na estratégia de maximização da diversidade (M). Os parâmetros genéticos avaliados para a coleção inteira e os modelos de coleção nuclear gerados foram: número de indivíduos (N_I), subamostras (N_S), total de alelos (N_A), conteúdo informativo de polimorfismo (PIC), heterozigosidade observada (H_O), heterozigosidade esperada (H_E) e Índice de Diversidade de Shannon (Sh). Foi possível caracterizar 17.412/2.370 PAVs/SNPs e 25.203 modelos gênicos com posições únicas no genoma. Houve uma correlação de Pearson positiva entre os modelos gênicos e o número de PAVs mapeados (R= 0,857; R^2= 0,734); e, do mesmo modo, com o número de marcadores SNPs (R= 0,708; R^2= 0,501). Foram obtidos modelos de coleção nuclear com 16, 26, 37, 55, 109, 127, 138 e 276 indivíduos. Devido ao bom ajuste dos parâmetros validados, tendo simultaneamente mantido o menor número de subamostras, o modelo MS3 (20% da coleção inteira) foi escolhido como o ideal para compor a coleção nuclear de caiaué. O conjunto de marcadores PAVs/SNPs mapeados proporciona uma cobertura consideravelmente homogênea ao longo do genoma e de regiões gênicas de *E. guineensis*. O modelo de coleção nuclear gerado neste trabalho irá permitir uma melhor utilização das subamostras na conservação genética de *E. oleifera*.

Palavras-chave: *Presence/Absence Variants*. SNPs. Diversidade genética. Germoplasma. Maximização da diversidade.

ABSTRACT

The objectives of this study were to identify and characterize PAVs (Presence / Absence Variants) and SNPs (Single Nucleotide Polymorphism) markers of American oil palm (*Elaeis oleifera*), based on the mapping of polymorphic sequences to African oil palm (*E. guineensis*) genome, and to outline core collections to *E. oleifera*. A bank of sequences generated by DArTSeq platform for 553 genotypes of *E. oleifera*was mapped against the public genome of African oil palm with BWA software. The SAMtools software package was used to identify SNPs. The gene models of date palm (*Phoenix dactylifera*) were mapped on the draft genome of *E. oleifera* developed by Embrapa. The genome was divided into intervals of 5 Mb (megabase) for a distribution analysis of genetic markers and gene models. From 1,666 SNPs were selected 500 with the high Shannon diversity index for the design of core collections from 553 individuals representing 206 accessions of *E. oleifera* based on maximizing strategy of diversity (M). The genetic parameters evaluated for the complete collection and the models were number of individuals (N_I), number of accessions (N_S), total number of alleles (N_A), polymorphism information content (PIC), observed heterozygosity (H_O), expected heterozygosity (H_E) and Shannon Diversity Index (Sh). It was possible to characterize 17,412/2,370 PAVs/SNPs and 25,203 gene models with unique positions on the genome. A Pearson positive correlation was found between the gene models and the number of mapped PAVs (R = 0.857, R2 = 0.734), and, likewise, with the number of SNP markers (R = 0.708, R2 = 0.501). Subsets with 16, 26, 37, 55, 109, 127, 138 and 276 accessions were selected. Because of the optimal adjustment of the validated parameters maintained while taking the least number of subsamples, the MS3 model (20% of entire collection) was chosen as the ideal to form the core collection of *E. oleifera*. The set of PAVs/SNPs markers mapped provide a uniform coverage throughout the genome and gene regions of *E. guineensis*. The core collection model generated in this work will allow better use of sub-samples in the genetic conservation of *E. oleifera*.

Keywords: Presence/Absence Variants. SNPs. Genetic diversity. Germplasm. Maximization of diversity.

1 INTRODUÇÃO

A palma de óleo africana (*Elaeis guineensis*) destaca-se por ser a oleaginosa cultivada com maior produtividade de óleo por área (SAVILAAKSO et al., 2014). Atualmente, o óleo de palma vem sendo utilizado também para a produção de biocombustíveis em decorrência da demanda por uma matriz energética mais sustentável, além de fatores relacionados ao aquecimento global, que abriram novas perspectivas para o uso de biodiesel (URQUIAGA; ALVES; BOODEY, 2005).

O caiaué ou palma de óleo americana (*E. oleifera*) apresenta produtividade inferior a das cultivares utilizadas em plantios comerciais de dendezeiro. Entretanto, o caiaué apresenta características vantajosas em relação ao dendezeiro, como menor taxa de crescimento do tronco, óleo mais insaturado e resistência/tolerância a diversas pragas e doenças (CUNHA et al., 2012).

O amarelecimento fatal (AF) é uma anomalia de etiologia desconhecida e uma grande ameaça para a expansão da dendeicultura no Brasil. Ainda não existe um método de controle eficaz contra o AF. O híbrido interespecífico BRS Manicoré, uma cultivar desenvolvida pela Embrapa a partir de cruzamentos entre *E. guineensis* e *E. oleifera*, foi testado em locais de incidência e não é afetado pelo AF (CUNHA et al., 2012).

Com a crescente demanda por óleo de palma, os programas de melhoramento genético da espécie vêm sendo forçados a obter ganhos em eficiência cada vez maiores. Nesse sentido, a aplicação de marcadores moleculares para fins de seleção assistida (SAM) vem se destacando como uma das principais abordagens do chamado melhoramento genômico.

Com o enorme avanço das tecnologias de sequenciamento de nova geração (NGS), têm-se hoje técnicas robustas que permitem a genotipagem de milhares de marcadores moleculares distribuídos ao longo de todo o genoma a

custos bem reduzidos. Estas técnicas são conhecidas como genotipagem por sequenciamento (GBS). De forma resumida, a metodologia de GBS ocorre mediante digestão do DNA genômico com enzimas de restrição e posterior indexação dos fragmentos com uma sequência de identificação. Esses fragmentos são então sequenciados por alguma plataforma de NGS, e as sequências geradas são preferencialmente mapeadas contra um genoma de referência para a chamada dos genótipos (com base em critérios de cobertura), gerando assim dados de marcadores PAVs (*presence/absence variants*). Através do alinhamento das sequências e identificação de variação de base única nestas, é possível se obter um conjunto de marcadores codominantes SNPs (*single nucleotide polymorphism*).

Uma aplicação importante de marcadores moleculares no melhoramento genético de plantas é na seleção de subamostras que melhor representam a diversidade genética em um banco de germoplasma, gerando assim o que é conhecido como coleção nuclear. Quando utilizados dados moleculares no desenvolvimento de coleção nuclear, a estratégia preferencialmente utilizada é a de maximização da diversidade (M). Com o desenvolvimento de uma coleção nuclear é possível diminuir o custo de manutenção do germoplasma e delimitar subamostras mais representativas de interesse no programa de melhoramento para fins de caracterização.

Diante do exposto, iniciou-se um projeto para o desenvolvimento de marcadores moleculares para a espécie americana de palma de óleo (*E. oleifera*), com base na plataforma DArTSeq. Os objetivos deste estudo foram, portanto, (i) identificar marcadores PAVs e SNPs, (ii) realizar a caracterização genômica desses marcadores, por meio da verificação de sua distribuição no genoma de *E. guineensis* e (iii) estabelecer um modelo de coleção nuclear para caiaué com marcadores SNPs.

2 MATERIAL E MÉTODOS

2.1 Material Vegetal

O tecido foliar foi coletado no Banco Ativo de Germoplasma (BAG) de caiaué mantido pela Embrapa Amazônia Ocidental (CPAA) no Campo Experimental do Rio Urubu - CERU, Rio Preto da Eva/AM em fevereiro de 2012 e armazenados a -80ºC. A extração de DNA ocorreu segundo o método CTAB modificado (DOYLE; DOYLE, 1990). A quantificação do DNA foi feita por meio de espectrofotometria (NanoDrop® ND-1000) e a integridade do DNA avaliada em géis de agarose 1%. As subamostras são originarias de diversos locais da bacia amazônica, compreendendo seis regiões do estado do Amazonas: Manaus, Rio Amazonas, Rio Solimões, Rio Negro, Caracarai e Rio Madeira. A população amostrada foi composta por 553 indivíduos (1 a 3 repetições por subamostra) de 206 subamostras (Apêndice B – Tabela 1).

2.2 Genotipagem por sequenciamento em plataforma DArTSeq para a construção da coleção nuclear

Uma vez extraído o DNA das subamostras, estes foram submetidos ao processo de redução de complexidade mediante digestão com enzimas de restrição (uma de corte frequente, *BstNI* e uma de corte raro, *PstI*) e marcados individualmente por meio de *barcodes* (ligados aos adaptadores PstI), e sequenciadas em conjunto usando a plataforma *Illumina HISeq* 2000. Após sequenciamento, os *reads* foram identificados e separados com base no *barcodes* em *pipeline* da empresa DArT Pty® (Austrália). Na ausência de um genoma de referência que pudesse ser utilizado para a chamada dos genótipos (*genotype call*), os *reads* únicos obtidos neste estudo foram utilizados para se montar uma

sequência de referência, a qual os *reads* individuais foram mapeados (*reads* praticamente idênticos foram combinados de modo que um ou mais SNPs no *read* não confundisse a análise). A partir desse mapeamento foi realizado o *score* dos *silico*-DArTs ou PAVs (*presence/absence variants*) (marcadores dominantes). Os marcadores resultantes foram então selecionados por meio da aplicação de filtros de qualidade baseados em *Call Rate* (>0,90), *Minor Allele Frequency* (MAF > 0,05) e *Q-value* (> 2).

2.3 Caracterização genômica de marcadores

Foi utilizado o banco de sequências geradas por meio da genotipagem por sequenciamento via plataforma DarTSeq de 553 indivíduos de caiaué. Foi realizado um pré-processamento para a remoção de sequências de baixa qualidade com os *softwares* FASTX-toolkit e FASTQC. As sequências foram mapeadas contra o genoma de referência de *E. guineensis* (SING et al., 2013). O mapeamento foi realizado utilizando o *software* BWA-MEM (parâmetros padrão) a partir da ferramenta de alinhamento Burrows-Wheeler (LI; DURBIN, 2010) para produzir um arquivo BAM (*Binary Alignment Map*).

Para a manipulação do arquivo, foi utilizado o pacote de *softwares* SAMtools (LI et al., 2009) (as opções utilizadas foram *sort*, *index*, *mpileupe* –uf e *bcftools view* –vcg) que também gerou o arquivo *Variant Call Format* (VCF) com os SNPs identificados nas sequências. Para ser considerado um marcador PAV, foi utilizado o filtro de cobertura de no mínimo 6X por genótipo e um máximo de dois SNPs na sequência. Um *script* em Perl (Apêndice C) foi utilizado para realizar a chamada dos genótipos para os PAVs identificados.

Os SNPs foram filtrados com base em parâmetros de qualidade *phred* 20, profundidade 10, qualidade de mapeamento 30 e qualidade consenso de 10. Posteriormente, cada um dos 16 cromossomos de *E. guineensis* foram divididos

em intervalos de 5Mb (megabase) para uma análise detalhada da distribuição dos marcadores ao longo do genoma.

Para observar a distribuição dos marcadores em relação às regiões gênicas do genoma de referência, foi realizado o mapeamento dos modelos gênicos públicos de *Phoenix dactylifera* (AL-MSSALLEM et al., 2013), esses modelos foram utilizados devido ao fato dos modelos gênicos para o gênero *Elaeis* spp. não estarem disponibilizados. Para o mapeamento dos genes, foi utilizado os *softwares nucmer* (– *maxmatch*), *delta-filter* (-*q*) e *show-coords* (-*rcl*) (KURTZ et al., 2004).

2.4 Marcadores SNPs

Um conjunto de 5.365 SNPs gerados pelo pipeline DArTPty® foi filtrado com base em *Call Rate* (>0,90) e MAF (>0,05). O número de marcadores foi reduzido para 1.666, e destes foram selecionados 500 com o maior índice de diversidade de Shannon gerado pela análise no *software* PowerCore (KIM et al., 2007). Esse conjunto de marcadores gerados foi alinhado contra o genoma de referência de *E. guineensis* por meio da ferramenta BLASTN (parâmetros padrão), os resultados foram filtrados aceitando somente sequências em regiões únicas no genoma e de no máximo dois *gaps* por sequência. O conjunto de 500 SNPs foi utilizado para o delineamento de coleções nucleares em caiaué.

2.5 Estabelecimento da coleção nuclear

O primeiro algoritmo foi baseado na estratégia M padrão que foi implementada pelo *software* MSTRAT (GOUESNARD et al., 2001). Para cada modelo gerado pelo MSTRAT foram utilizadas 100 replicações independentes e

100 interações, a composição do modelo escolhido foi selecionada com base no maior índice de diversidade de Shannon. O segundo algoritmo foi baseado em um modelo heurístico implementado pelo *software* PowerCore (KIM et al., 2007).

2.6 Análise de diversidade genética na coleção inteira

Os parâmetros genéticos avaliados para a coleção inteira de caiaué foram: número total de alelos (N_A), conteúdo informativo de polimorfismo (PIC), heterozigosidade observada (H_O), heterozigosidade esperada (H_E) e índice de diversidade de Shannon (Sh). O *software* PowerMaker v3.25 (LIU; MUSE, 2005) foi utilizado nas análises.

2.7 Validação dos modelos de coleção nuclear de caiaué

Os parâmetros genéticos avaliados para a coleção inteira (N_A, PIC, H_E, H_O e Sh) foram estimados separadamente para cada um dos modelos obtidos, o *software* utilizado nessa análise foi o PowerMaker v3.25 (LIU; MUSE, 2005).

A estruturação genética foi avaliada por meio do *software* GenAlEX 6.1 (PEAKALL; SMOUSE, 2006) com a Análise de Componentes Principais (PCA) a partir de uma matriz de distância genética de Nei (1978) para os 553 indivíduos da coleção inteira de caiaué em relação a cada modelo gerado de coleção nuclear. Análise da Variância Molecular (AMOVA) foi obtida para comparar a diversidade entre os modelos de coleção nuclear e dentro dos modelos. Essa análise foi obtida utilizando o *software* GenAlEx 6.1 (PEAKALL; SMOUSE, 2006).

3 RESULTADOS

3.1 Caracterização genômica de marcadores PAVs e SNPs

Um total de 849.301.904 sequências foram mapeadas contra o genoma de referência de *E. guineensis*. Destas, foi possível caracterizar 17.412 PAVs com posições únicas no genoma. Com base em um filtro de no mínimo 6X, foi realizada a chamada dos PAVs identificados para os 553 indivíduos, após aplicar filtros de *Call Rate* (>0,90) e MAF (>0,05) foram mantidos 2.423 marcadores.

Com base no alinhamento dessas sequências, foram posteriormente identificados 36.590 SNPs, sendo 5.433 do tipo heterozigoto. Após aplicar os parâmetros de filtragem, esse número foi reduzido para 2.370 SNPs (*high-quality* SNPs). Devido à baixa cobertura de base SNP para alguns genótipos (média de 0,37 para os 553 genótipos), não foi possível pelo método de GBS, fazer a chamada por genótipo com confiabilidade desses SNPs identificados.

Foi possível mapear contra o genoma de referência um total de 25.203 modelos gênicos de *P. dactilylifera*. O número de marcadores PAVs/SNPs nos 16 cromossomos variou de 569 a 2.228 e 66 a 322 respectivamente. O número médio de marcadores PAVs/SNPs e modelos gênicos por cromossomo foi de 1.088, 148 e 1.575 respectivamente. Para os 130 intervalos em que os cromossomos foram divididos, houve uma frequência de PAVs/SNPs e modelos gênicos de no mínimo 33, 1 e 7 e no máximo 352, 46 e 605, com uma média de 134, 18 e 194 por intervalo respectivamente (Figura 1).

A média de distância física entre PAVs adjacentes foi de 38,7 kb enquanto que para os marcadores SNPs foi de 280,9 kb. Foi observado que um maior número de PAVs/SNPs (59%) está a uma distância igual ou maior de 20 kb em relação ao modelo gênico mais próximo (Figura 2). A relação entre o número de marcadores e o de modelos gênicos foi evidenciada por uma

correlação de Pearson positiva do número de PAVs mapeados (R= 0,857; R^2= 0,734), e, do mesmo modo, com o número de marcadores SNPs (R= 0,708; R^2= 0,501) (Figura 3 e 4).

3.2 Avaliação genômica dos marcadores, estabelecimento de coleção nuclear e comparação com a coleção inteira

Dos 500 marcadores SNPs utilizados neste trabalho, foi possível mapear 245 (49%) contra o genoma de *E. guineensis*. Estes foram distribuídos ao longo de todos os 16 cromossomos do genoma. O número de marcadores variou de 5 a 31 SNPs, para o cromossomo 1 e 15 respectivamente (Figura 5).

Foram avaliados seis modelos de coleção nuclear com 6, 10, 20, 23, 25 e 50% (37, 55, 109, 127, 138, 276 indivíduos respectivamente), nomeados MS1, MS2, MS3, MS4, MS5 e MS6, respectivamente, esses modelos foram gerados pelo *software* MSTRAT. A coleção nuclear estabelecida por meio da pesquisa aleatória do *software* PowerCore foi composta por 26 indivíduos (4,7%), e aquela realizada por meio do algoritmo heurístico foi composta por 16 indivíduos (2,8%), nomeados PC1 e PC2 respectivamente.

Todos os alelos detectados na coleção inteira de *E. oleifera* foram mantidos nos diferentes modelos de coleção nuclear (Tabela 1). Um maior valor de PIC em relação a CI (coleção inteira) foi observado somente no modelo MS6, já os modelos MS4 e MS5 apresentaram o mesmo valor que a CI e o restante dos modelos apresentaram um valor menor.

O valor de H_O foi maior que o observado para a CI (0,271) em todos os modelos, sendo o modelo PC1 o com maior valor (0,314) e MS5 com o menor (0,273). Em contraste, o valor de H_E só foi maior que o apresentado pela CI (0,463) para o modelo MS6 (0,464) e o modelo MS4 apresentou o mesmo valor (0,463). Foi observado que os valores de H_E são superiores aos de H_O (Tabela 1).

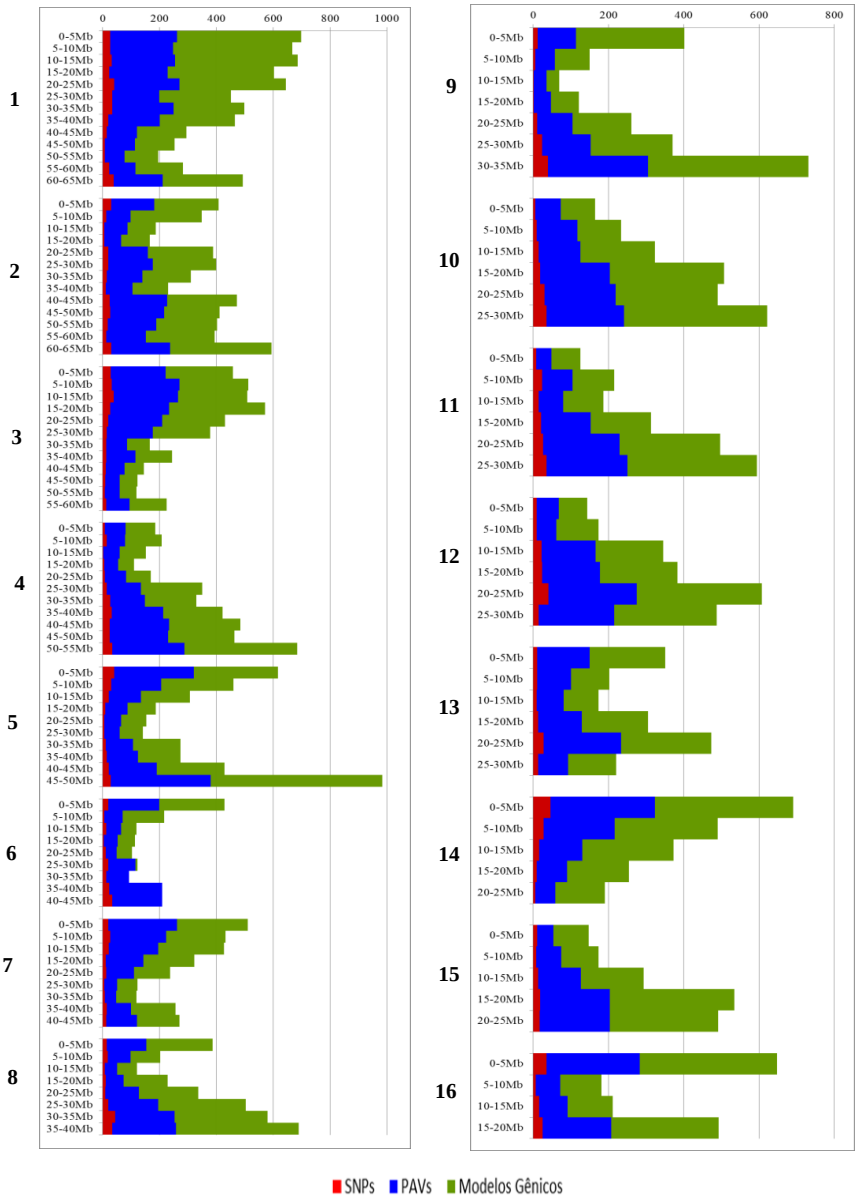

Figura 1 Análise da frequência de marcadores PAVs/SNPs e modelos gênicos no genoma de *Elaeis guineensis*

Legenda: Os 16 cromossomos foram divididos em 130 intervalos de 5Mb (megabase). Em vermelho a frequência de SNPs, em azul de PAVs e em verde de modelos gênicos.

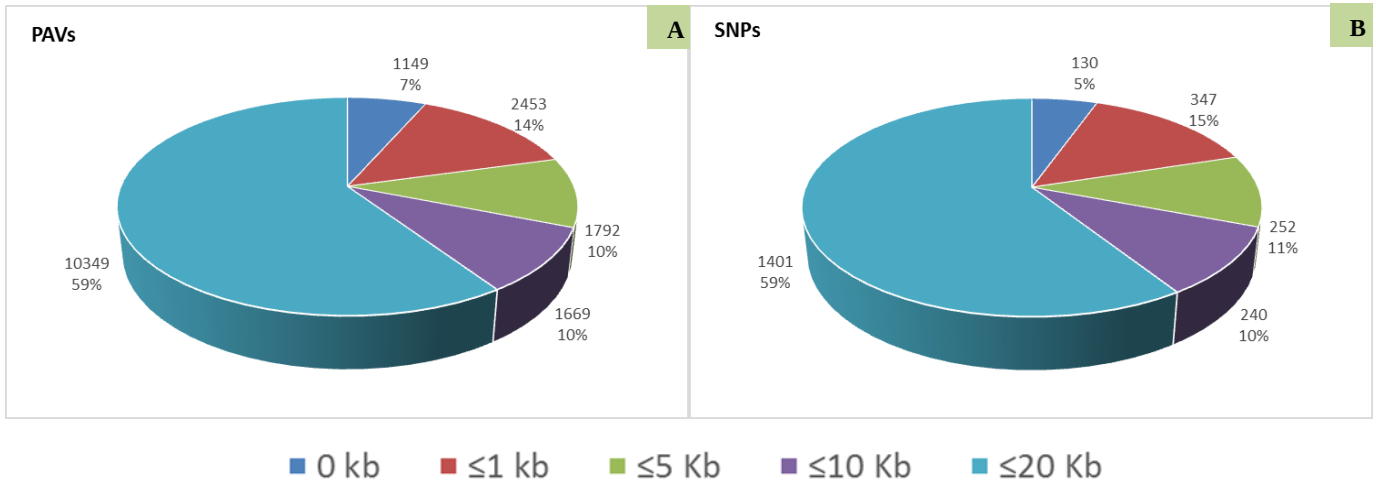

Figura 2 Distribuição de marcadores em relação ao modelo gênico mais próximo

Legenda: A) Distância em kilobase (kb) de marcadores PAVs em relação ao modelo gênico mais próximo; B) Distância em kilobases (kb) de marcadores SNPs em relação ao modelo gênico mais próximo.

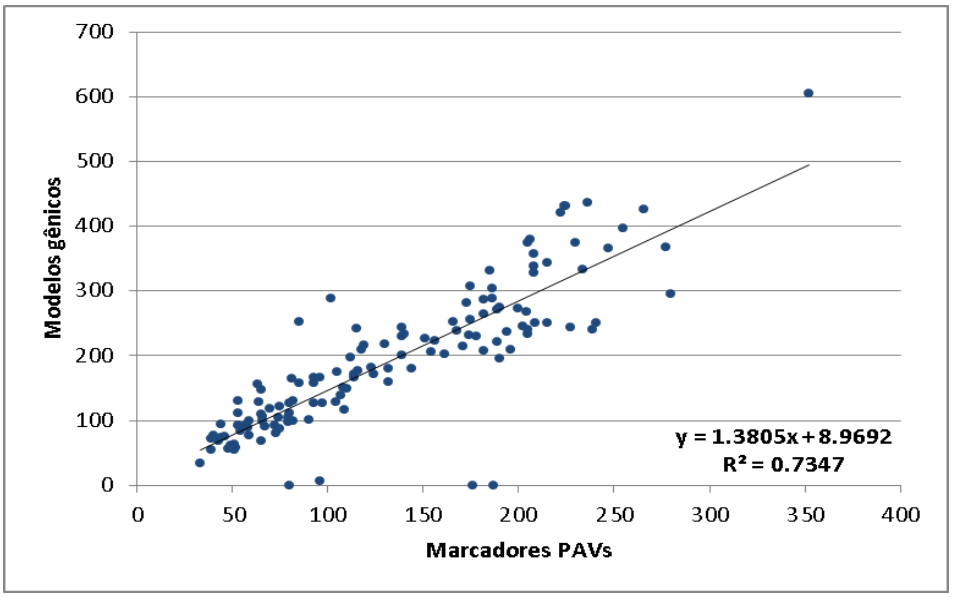

Figura 3 Correlação entre o número de marcadores PAVs e os modelos gênicos mapeados no genoma de *Elaeis guineensis*

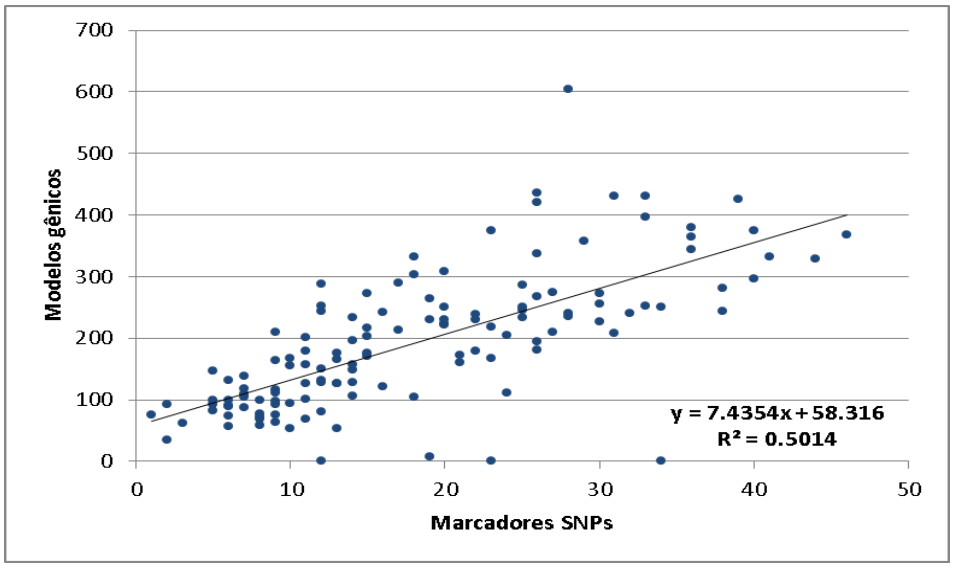

Figura 4 Correlação entre o número de marcadores SNPs e os modelos gênicos mapeados no genoma de *Elaeis guineensis*

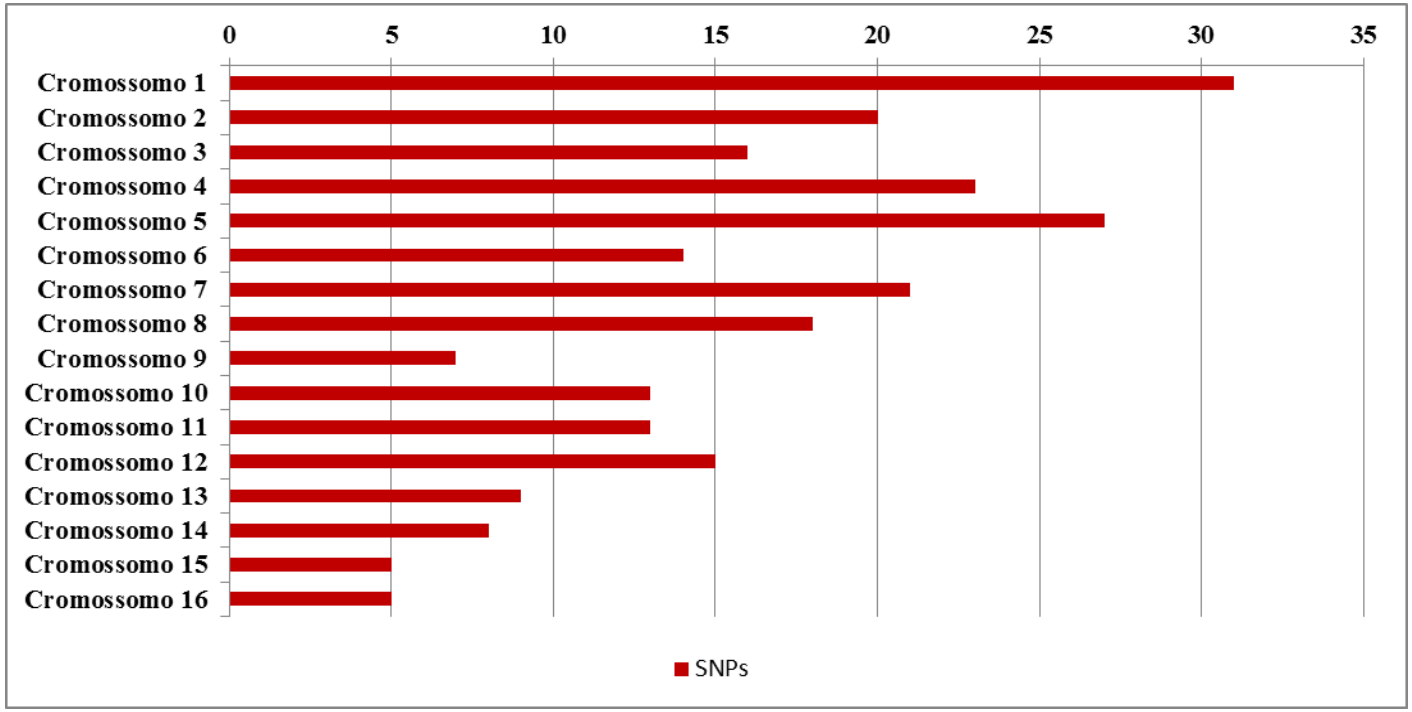

Figura 5 Distribuição de 500 marcadores SNPs gerado pelo pipeline da DArT Pty utilizados para geração de coleção nuclear de *Elaeis oleifera* ao longo dos 16 cromossomos do genoma de *E. guineensis*

Tabela 1 Parâmetros genéticos avaliados para a coleção inteira e para os diferentes modelos de coleção nuclear de *Elaeis oleifera*

Parâmetros	CI	PC1	PC2	MS1	MS2	MS3	MS4	MS5	MS6
N_I	553	16	26	37	55	109	127	138	276
N_S	206	16	25	35	52	86	105	108	197
N_A	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PIC	0,355	0,351	0,351	0,352	0,352	0,354	0,355	0,355	0,356
H_O	0,271	0,314	0,289	0,275	0,277	0,274	0,280	0,273	0,282
H_E	0,463	0,457	0,456	0,458	0,459	0,461	0,463	0,462	0,464
Sh	0,693	0,677	0,683	0,687	0,689	0,691	0,691	0,691	0,692

N_I = número de indivíduos; N_S = número de subamostras N_A = número total de alelos; PIC = conteúdo informativo de polimorfismo; H_O = heterozigosidade observada; H_E = heterozigosidade esperada; Sh=Índice de Diversidade de Shanoon.

O índice de diversidade de Shannon (Sh) para a coleção inteira foi de 0,693. Para os modelos de coleção nuclear, o índice variou de 0,692 a 0,677, para MS6 e PC1 respectivamente. Ocorreu um aumento do índice em relação ao número de subamostras.

Em relação às regiões de coleta do material da coleção inteira (Manaus, Rio Amazonas, Rio Solimões, Rio Negro, Caracarai e Rio Madeira) todas as coleções mantiveram pelo menos um indivíduo de cada região (Figura 6). A região mais representada na coleção inteira foi à do Rio Madeira e a de menor representação foi a da Região Caracarai, essa mesma proporção foi mantida nos modelos MS3, PC2 e MS6.

Não houve nenhum indivíduo comum para todos os oito modelos de coleção nuclear. Quatro indivíduos foram comuns em sete modelos. No entanto, muitos indivíduos (209) estão representados apenas em um dos modelos. Somente o modelo PC1 apresentou um indivíduo por subamostra, para os demais, o número de subamostras com mais de um indivíduo variou de 1 a 79.

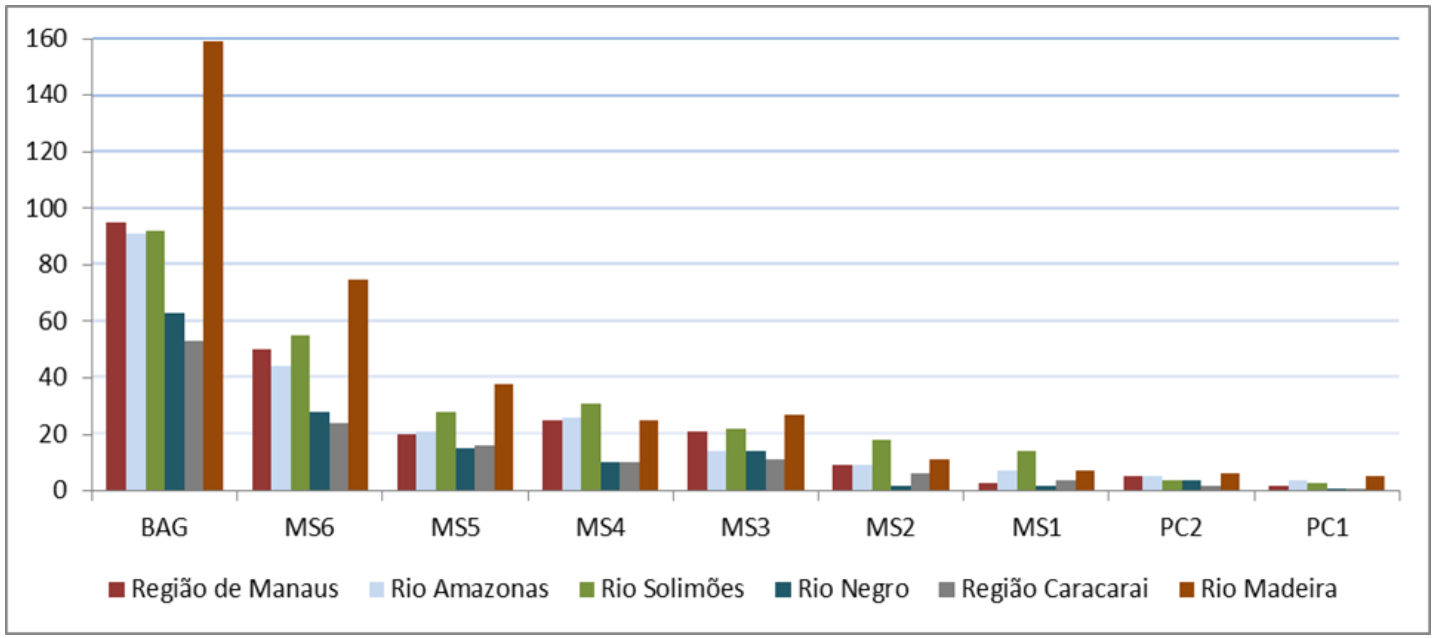

Figura 6 Histograma de distribuição das subamostras da coleção inteira (553 indivíduos) e para modelos de coleção nuclear PC1 (16 indivíduos), PC2 (26 indivíduos), MS1 (37 indivíduos), MS2 (55 indivíduos), MS3 (109 indivíduos), MS4 (127 indivíduos), MS5 (138 indivíduos) e MS6(276 indivíduos) de *Elaeis oleifera* em relação às regiões de coleta

3.3 Diferenciação dos modelos de coleção nuclear

Análise de variância molecular (AMOVA) foi utilizada para avaliar a variação genética dentro e entre os modelos de coleção nuclear de caiaué e mostrou que toda a variação molecular (100%) se encontra dentro dos modelos, ficando a variação entre os modelos em zero (Tabela 2). A porcentagem de variação molecular, quando considerado cada modelo, variou de 1,96 a 35,33%. Os modelos MS5 e MS6 foram os que apresentaram maior diferenciação genética comparados com os outros modelos, com 17,63 e 35,33% respectivamente.

Tabela 2 Análise de variância molecular (AMOVA) entre os diferentes modelos de coleção nuclear de caiaué (*Elaeis oleifera*)

Fonte de Variação	Grau de Liberdade	Soma de quadrados	Quadrados médios	Porcentagem de Variação (%)
Entre os modelos	7	1.703.687	243,384	0,0
Dentro dos modelos	776	194.324.048	250,418	100,00
PC1	15	3.812.438	254,163	1,96
PC2	25	6.286.346	251,454	3,23
MS1	36	9.004.324	250,120	4,63
MS2	54	13.519.527	250,362	6,96
MS3	108	27.052.679	250,488	13,92
MS4	126	31.729.339	251,820	16,33
MS5	137	34.262.884	250,094	17,63
MS6	275	68.656.511	249,660	35,33
Total	783	196.027.735	-	100,00

3.4 Validação dos modelos de coleção nuclear com análise de componentes principais - PCA

Os resultados de PCA mostraram que em geral todos os modelos de coleção nuclear foram representativos em relação à distribuição da coleção

inteira (Figura 7). Para os diferentes agrupamentos ocorreu em média uma explicação de 46,17% para a variação genética. O modelo MS6 apresenta uma alta representatividade de subamostras em relação à coleção inteira. Em contraste, modelos com um menor número de subamostras (PC1, PC2 e MS1) conseguiram representar de forma satisfatória a distribuição da coleção inteira, com subamostras distribuídas em todos os quadrantes da representação de PCA.

4 DISCUSSÃO

Este estudo fornece dados inéditos de genotipagem por sequenciamento (GBS) via plataforma DArTseq para uma espécie de palmeira, e é o primeiro a investigar as propriedades desse tipo de marcador em relação a um mapa físico do genoma e de modelos gênicos. Além disso, é o primeiro relato de desenvolvimento de coleção nuclear para caiaué e de utilizar marcadores SNPs baseados em NGS para esse fim.

4.1 Caracterização genômica de marcadores

Os resultados encontrados neste trabalho indicam que o número total de marcadores PAVs/SNPs de *E. oleifera* mapeados contra o genoma de referência tende a fornecer marcadores em essencialmente todos os intervalos em que o genoma foi dividido e uma forte correlação positiva (R= 0,857 e R= 0,708 para PAVs/SNPs respectivamente) no aumento do número de marcadores em relação aos modelos gênicos que foram também mapeados contra o genoma.

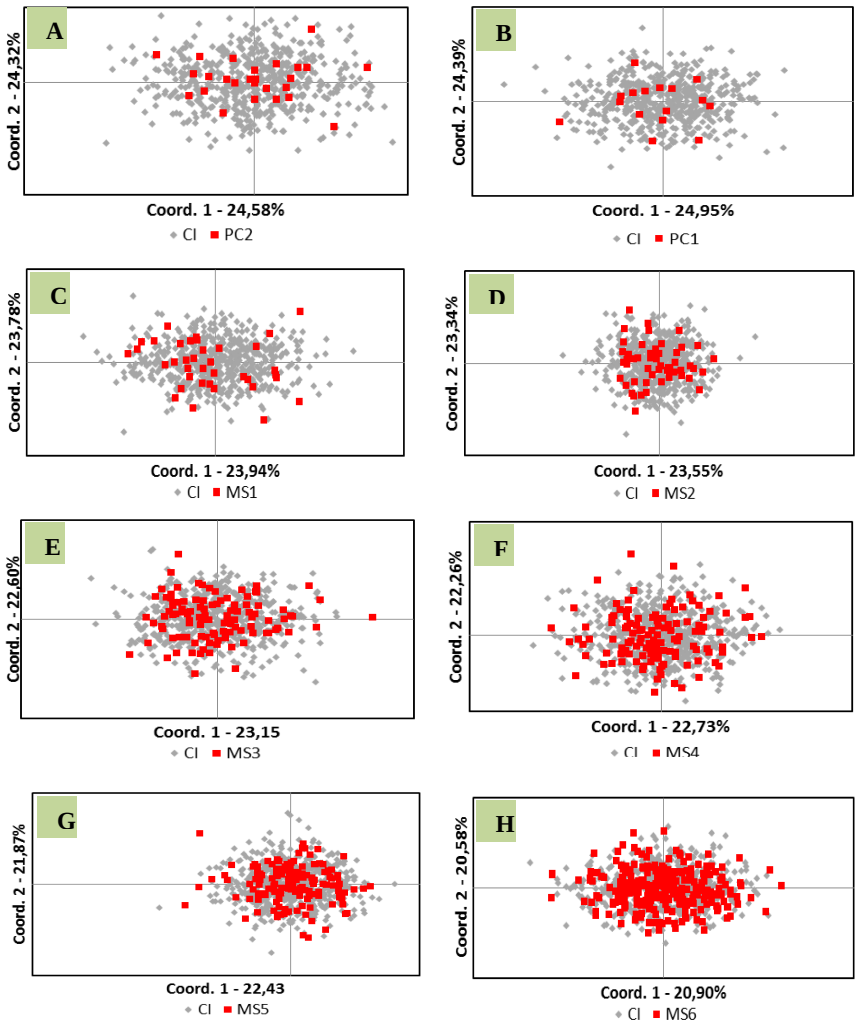

Figura 7 Análise de Componentes Principais (PCA) de 553 indivíduos da coleção inteira em relação cada modelo de coleção nuclear de caiaué (*Elaeis oleifera*)

Legenda: A) Coleção inteira e PC1 com 16 indivíduos (2,8%); B) Coleção inteira e PC2 com 26 indivíduos (4,7%) C) Coleção inteira e MS1 com 37 indivíduos (6%); D) Coleção inteira e MS2 com 55 indivíduos (10%); E) Coleção inteira e MS3 com 109 indivíduos (20%); F) Coleção inteira e MS4 com 127 indivíduos (23%); G) Coleção inteira e MS5 com 138 indivíduos (25%); H) Coleção inteira e MS6 com 276 indivíduos (50%).

Ainda existem poucos trabalhos sobre caracterização genômica de marcadores gerados pela tecnologia DArT e por GBS. Petroli et al. (2012) alinhou 6.571 sondas DArT em conjunto com a distribuição de 41.204 modelos gênicos no genoma do *Eucalyptus* e encontrou cobertura homogênea dos marcadores ao longo do genoma com uma amostragem preferencial de regiões gênicas. Em soja, com a utilização de GBS, foram identificados 16.502 locos SNPs em 301 cultivares, sendo 68,75% alinhados em posições únicas no genoma de referência (JARQUÍN et al., 2014).

O mapa físico da montagem do genoma de *E. guineensis* utilizado neste trabalho foi realizada com 26 coberturas de *reads* da tecnologia Roche/454 ancorado em um mapa genético densamente saturado com marcadores microssatélites, SNPs e RFLP (SING et al., 2013), o que justifica a utilização desse genoma para a caracterização dos marcadores de caiaué. Já os modelos gênicos públicos que foram mapeados nesse genoma são de *Phoenix dactylifera* que é uma espécie de palmeira relacionada filogeneticamente ao gênero *Elaeis* spp. (SING et al., 2013).

A relação dos marcadores mapeados em associação com os modelos gênicos corroboram outros estudos que relatam que marcadores baseados em endonucleases estão predominantemente localizados em regiões do genoma de baixa cópia e ricas em genes (PETROLI et al., 2012; WENZL et al., 2004). Regiões gênicas são preferencialmente selecionadas devido à afinidade das enzimas de restrição utilizadas em protocolos GBS na digestão em regiões de metilação no genoma (SCHNABLE et al., 2009).

Neste trabalho, não foi possível realizar com confiabilidade a chamada por genótipos para os SNPs identificados devido à baixa cobertura destes para diferentes genótipos. Esse resultado indica que a cobertura amostrada com as técnicas de GBS é desigual ao longo do genoma. Beissinger et al. (2013) investigaram a densidade de marcadores e a profundidade de *reads* gerados por

GBS em milho, e relataram que a cobertura para diferentes locais do genoma é extremamente variável, aproximadamente 76% de potenciais sítios de restrição não foram amostradas pela técnica, enquanto outras regiões possuíam cobertura de até 2.369 vezes a média esperada.

A técnica de DArTSeq permitiu a genotipagem de um painel relativamente grande de marcadores em uma espécie que até então possuía poucos recursos genômicos. Os marcadores têm distribuição relativamente homogênea no genoma, amostrando de forma robusta as regiões gênicas que foram analisadas. Os marcadores SNPs identificados que não puderam ser genotipados via GBS podem ser utilizados em outras plataformas de genotipagem desse tipo de marcador.

As informações relativas à distribuição física dos marcadores no genoma permitiram a seleção de um conjunto específico de marcadores para o desenho de sondas, esse é um método de GBS de regiões específicas do genoma-alvo, em que as sondas são selecionadas para hibridar com regiões únicas de interesse no genoma de referência. Essa iniciativa representa um grande avanço no estudo genômico de caiaué.

A seleção genômica possui vantagem sobre os métodos de seleção tradicional devido principalmente à economia no tempo de seleção, principalmente para espécies perenes como a palma de óleo que possui um ciclo reprodutivo longo (ALVES et al., 2014). Além disso, o conjunto específico de marcadores relacionados a regiões gênicas podem ser usados em estudos de filogenia, genética de populações, seleção assistida por marcadores e seleção genômica ampla.

4.2 Estabelecimento de coleção nuclear em caiaué

Para os parâmetros genéticos avaliados houve pouca alteração dos apresentados pelos modelos de coleções nucleares em relação àqueles obtidos para a coleção inteira. Além disso, todos os modelos mantiveram subamostras representando as seis regiões do Estado do Amazonas em que a coleção inteira está dividida.

Neste estudo, 100% dos alelos SNPs apresentados pela coleção inteira foram mantidos em todos os modelos. Evidências sugerem que a estratégia M quando utilizada com marcadores moleculares é capaz de manter a diversidade genética e alélica para as diferentes coleções nucleares geradas (MARITA; RODRIGUEZ; NIENHUIS, 2000; SCHOEN; BROWN, 1993). No entanto, manter o número total de alelos em coleções nucleares nem sempre é possível dependendo da espécie ou do conjunto de dados; em trigo somente 98% dos alelos foram mantidos na coleção nuclear (BALFOURIER et al., 2007). Belaj et al. (2012), utilizando marcadores microssatélites para construção de coleção nuclear em *Olea europaea*, obtiveram uma redução em média de 10% na quantidade total de alelos da coleção inteira para os cinco modelos de coleção nuclear que foram gerados.

Os valores de H_E foram superiores daqueles obtidos para o índice de H_O, o que pode estar relacionado a um maior número de homozigotos na população devido a algum efeito populacional de endogamia ou pode ser resultante de um efeito amostral da coleção inteira.

Os parâmetros genéticos avaliados não variaram de forma drástica para os modelos gerados pelos métodos implementados pelos *softwares* PowerCore e MSTRAT. A estratégia M (SHOEN; BROWN, 1993) é capaz de selecionar de forma eficiente subamostras que maximizam a diversidade genética apresentada pela coleção inteira. A implementação dessa estratégia pelo *software* MSTRAT

(GOUESNARD et al., 2001) permite que os alelos sejam selecionados de uma forma interativa, mantendo a diversidade por critérios de riqueza alélica na análise. Em contraste, PowerCore (KIM et al., 2007) utiliza uma busca por meio de um algoritmo heurístico que possui a capacidade de representar todos os alelos identificados por marcadores moleculares e todas as classes de observações fenotípicas no desenvolvimento de coleções nucleares.

O resultado da AMOVA, em que toda a variação se encontra dentro dos modelos, já era esperado devido à composição genética próxima entre os modelos. Oliveira et al. (2014) também obtiveram resultados semelhantes aos encontrados neste trabalho, com 0,30 e 99,70% da variação entre e dentro dos modelos respectivamente.

A análise de PCA é um método exploratório que auxilia na elaboração de hipóteses mais concretas a partir dos dados coletados (MINGOTI, 2005). A análise de PCA apresenta indícios que estão relacionados com o nível de explicação para as relações genéticas existentes. De acordo com os resultados da análise de PCA, observa-se que para todos os modelos plotados contra a coleção inteira, a maioria dos indivíduos está dispersa no painel, representando subamostras nos quatro quadrantes da dispersão gráfica. Em média, 46% da variância total foi explicada nas duas dimensões analisadas, no entanto, para cada agrupamento esse valor foi reduzido em função do aumento no número de subamostras.

Em geral, uma coleção nuclear pode ter 10% do tamanho original o que representa aproximadamente 70% da diversidade da coleção original (BROWN, 1989). No entanto, diferentes porcentagens podem ser aceitas devido à estabilização da coleção nuclear depender de diversos fatores, tais como, o tamanho da coleção original, a qualidade dos dados coletados para a caracterização e avaliação da estratificação da coleção original e a estratégia de amostragem utilizada (COCHRAN, 1977). Uma boa coleção nuclear incorpora o

máximo da diversidade da espécie com o mínimo de redundância no menor tamanho possível para facilitar a manipulação da coleção (BROWN, 1989).

O BAG de caiaué da Embrapa possui aproximadamente 4.000 plantas, conservadas em 30 hectares de plantio, no Campo Experimental do Rio Urubu (CERU). Apesar da alta diversidade genética presente no BAG de caiaué (MORETZSOHN et al., 2002), a sua exploração efetiva é baixa, uma vez que apenas 20% das subamostras mantidas no BAG foram caracterizadas e grande parte da caracterização e avaliação já não são mais possíveis de serem realizadas considerando a idade elevada das plantas.

A criação da coleção nuclear do caiaué tem relevância tanto para o desenvolvimento científico e tecnológico do estado do Amazonas e da região Norte, quanto em nível nacional, com a expansão da cultura no Brasil. O delineamento de coleção nuclear para essa espécie também reduziria os custo de estudos realizados no campo experimental por delimitar as subamostras mais representativas da coleção inteira.

Somente o modelo PC1 não apresentou mais de um indivíduo por subamostra, possivelmente devido ao tamanho reduzido desse modelo (16 subamostras), para os demais modelos, a variação de número de subamostracom mais de um indivíduo foi de 1 a 79. Esse resultado pode indicar que existe considerável variabilidade genética dentro de subamostras de uma mesma população. Em um estudo de diversidade genética realizado na Embrapa Agroenergia com PAVs para o banco de germoplasma de caiaué, foi encontrado maior variabilidade dentro de subamostras (56,36%) e quando analisado o agrupamento de 206 subamostras com três replicações, foi observado uma discordância de 13,20% (Dados não publicados).

O modelo MS3 apresentou um índice de diversidade de Shannon de 0,691 mantendo-se muito próximo do avaliado para a coleção inteira (0,693), os demais parâmetros genéticos avaliados também apresentaram pouca variação,

dessa forma esse modelo que reteve 20% do total de indivíduos da coleção inteira é a escolha mais apropriada como coleção nuclear de *E. oleifera*. Por maximizar a diversidade genética e reduzir o número de genótipos, a coleção MS3 pode facilitar estudos de variabilidade e correlação com características morfológicas de interesse agronômico no programa de melhoramento genético de *E. oleifera* desenvolvido pela Embrapa.

A proposta de desenvolvimento neste estudo pode ser eficientemente aplicada para o desenvolvimento de coleção nuclear em *E. oleifera* e outras espécies. No entanto, um conjunto ideal de subamostras de uma coleção nuclear deve ser constantemente revisado quando novas subamostras são obtidas ou quando novas informações biológicas são coletadas. Outros estudos com diferentes espécies (MCKHANN et al., 2004; BALFOURIER et al., 2007; BELAJ et al., 2012) indicam que a diversidade genética de uma coleção nuclear pode ser mais bem alcançada quando dados moleculares são usados em conjunto com características fenotípicas.

5 CONCLUSÕES

O conjunto de marcadores PAVs/SNPs de *Elaeis oleifera* mapeados neste estudo proporciona uma cobertura consideravelmente homogênea ao longo do genoma de referência possuindo marcadores em todos os intervalos que possuem modelos gênicos. Estes marcadores poderão agora ser aplicados em estudos de diversidade genética e em estratégias de SAM.

O modelo de coleção nuclear MS3 (20% de indivíduos em relação à coleção inteira) desenvolvido neste trabalho irá facilitar futuros trabalhos de melhoramento e conservação genética de *E. oleifera* por manter a diversidade genética com um menor número de subamostras.

REFERÊNCIAS

ALVES, A. A. et al. Perennial plants for biofuel production: bridging genomics and field research. **Biotechnology Journal**, Amsterdam, v. 10, n. 1, p. 2-4, Nov. 2014.

AL-MSSALLENS, I. S. et al. Genome sequence of the data palm *Phoenix dactylifera* L. **Nature Communications**, London, v. 4, n. 2274, p. 1-9, Aug. 2013.

BALFOURIER, F. et al. A worldwide bread wheat core collection arrayed in a 384-well plate. **Theoretical and applied genetics**, Berlin, v. 114, n. 7, p. 1265-1275, May 2007.

BEISSINGER, T. M. et al. Marker density and read depth for genotyping populations using genotyping-by-sequencing. **Genetics**, Austin, v. 193, n. 4, p. 1073-1081, Apr. 2013.

BELAJ, A. et al. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. **Tree Genetics & Genomes**, Amsterdam, v. 8, n. 1, p. 365-378, Dec. 2012.

BROWN, A. H. D. Core collections: a practical approach to genetic resources management. **Genome**, Ottawa, v. 31, n. 2, p. 818-824, Jan. 1989.

COCHRAN, W. G. **Sampling techniques**. 3. ed. New York: John Wiley & Sons, 1977. 448 p.

CUNHA, R. N. V. et al. Domestication and breeding of the American Oil Palm. In: BORÉM, A.; LOPES, M. T. G.; CLEMENT, C. R. (Ed.). **Domestication and breeding:** Amazon species. Viçosa: Suprema, 2012. p. 275-296.

CUNHA, R. N. V.; LOPES, R. **BRS Manicoré:** híbrido interespecífico entre o caiaué e o africano recomendado para áreas de incidências de amarelecimento-fatal. Manaus: Embrapa Amazônia Ocidental, 2010. 3 p.

DOYLE, J. J.; DOYLE, J. L. Isolation of plant DNA from fresh tissue. **Focus**, Rockville, v. 12, n. 8, p. 13-15, Dec. 1990.

GOUESNARD, B. et al. MSTRAT: an algorithm for building germ plasm core collection by maximizing allelic or phenotypic richness. **Journal Heredity**, Washington, v. 92, n. 1, p. 93-94, Jan./Feb. 2001.

JARQUÍN, D. et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. **BMC Genomics**, London, v. 15, n. 740, p. 1-10, Aug. 2014.

KIM, K. W. et al. PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. **Bioinformatics**, Oxford, v. 23, n. 16, p. 2155-2162, Aug. 2007.

KURTZ, S. et al. Versatile and open software for comparing large genomes. **Genome Biology**, London, v. 5, n. 2, p. 1-12, Jan. 2004.

LI, H.; DURBIN, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. **Bioinformatics**, Oxford, v. 26, n. 5, p. 589-595, Mar. 2010.

LI, H. et al. 1000 genome project data processing subgroup. The sequence alignment/map format SAMtools. **Bioinformatics**, Oxford, v. 25, n. 16, p. 2078-2079, Aug. 2009.

LIU, K.; MUSE, S. V. PowerMarker: an integrated analysis environment for genetic marker analysis. **Bioinformatics**, Oxford, v. 21, n. 9, p. 2128-2129, May 2005.

MARITA, J. M.; RODRIGUEZ, J. M.; NIENHUIS, J. Development of an algorithm identifying maximally diverse core collections. **Genetic Resources and Crop Evolution**, Dordrecht, v. 47, n. 5, p. 515-526, Oct. 2000.

MCKHANN, H. I. et al. Nested core collections maximizing genetic diversity in Arabidopsis thaliana. **Plant Journal**, Malden, v. 38, n. 1, p. 193-202, Apr. 2004.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Editora da UFMG, 2005. 295 p.

MORETZSOHN, M. C. et al. Genetic diversity of Brasilian oil palm (*Elaeis oleifera* H. B. K.) germplasm collected in the Amazon forest. **Euphytica**, Wageningen, v. 124, n. 1, p. 35-45, May 2002.

OLIVEIRA, E. J. et al. Development of a cassava core collection based on single nucleotide polymorphism markers. **Genetics and Molecular Research**, Ribeirão Preto, v. 13, n. 3, p. 6472-6485, Aug. 2014.

PEAKALL, R.; SMOUSE, P. E. GENALEX 6: genetic analysis in Excel. Population genetic software for teching and research. **Molecular Ecology Notes**, Oxford, v. 6, n. 1, p. 288-295, Mar. 2006.

PETROLI, C. D. et al. Genomic characterization of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. **PLoS One**, San Francisco, v. 7, n. 9, p. 1-14, Sept. 2012.

SAVILAAKSO, S. et al. Systematic review of effects on biodiversity from oil palm production. **Environmental Evidence**, London, v. 3, n. 4, p. 1-20, Feb. 2014.

SCHNABLE, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. **Science**, Washington, v. 326, n. 5956, p. 1112-1115, Nov. 2009.

SCHOEN, D. J.; BROWN, A. H. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 90, n. 22, p. 10623-10627, Nov. 1993.

SING, R. et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. **Nature**, London, v. 500, n. 7462, p. 335-339, July 2013.

URQUIAGA, S.; ALVES, B. J. R.; BOODEY, R. M. Produção de biocombustíveis: a questão do balanço energético. **Revista de Política Agrícola**, Brasília, v. 14, n. 1, p. 42-46, Jan. 2005.

WENZL, P. et al. Diversity arrays technology (DArT) for whole-genome profiling of barley. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 15, n. 740, p. 9915-9920, June 2004.

CONSIDERAÇÕES FINAIS

O estudo da genética e da genômica de caiaué são de extrema importância na busca por características de interesse a serem incorporadas no melhoramento genético desta espécie, como também no desenvolvimento de híbridos interespecíficos com *E. guineensis*.

Diante das descobertas descritas neste estudo, espera-se avançar nos trabalhos de pesquisa com essa espécie, principalmente no que diz respeito ao aperfeiçoamento do *draft* do genoma de caiaué com uma montagem mais representativa, identificação e caracterização funcional das regiões gênicas, utilização dos marcadores identificados em estratégia de seleção assistida por marcadores e seleção genômica ampla, além de estudos de diversidade genética e melhoramento.

O modelo de coleção nuclear que foi estabelecido para caiaué irá contribuir na conservação e caracterização desse banco de germoplasma, permitindo o direcionamento de estudos genéticos realizados com a espécie para os indivíduos mais representativos do banco. Outras informações moleculares e agronômicas podem ser agregadas a esse estudo com o propósito de aperfeiçoar o modelo gerado.

APÊNDICE

APÊNDICE A – *Script* em *Perl* desenvolvido para a análise do resultado de coordenadas de alinhamento gerado pelo *software* Nucmer.

```
#!/usr/bin/perl –w

# Quick script to parse mummer/nucmer output .coords and calculate
reference coverage

#

# Edu :) 20140606

#

# Usage: perl parse.coords.cov.pl <filename.coords> [output name]

#

# Expected pattern

#
/lbb/analise_temp/prodende/map/z/E_oXDraft_M/Pub_Eo_Scaf_KE504633-
KE519432.fasta
/lbb/analise_temp/prodende/map/z/E_oXDraft_M/ManicoDraft1_Edu.scaf.fa
sta

# NUCMER

#

#    [S1]     [E1] |     [S2]     [E2] | [LEN 1]  [LEN 2] | [% IDY] | [LEN R]
[LEN Q] | [COV R]  [COV Q] | [TAGS]

#==================================================
==================================================
====================

#      26     1730 |     4245     2525 |     1705     1721 |    96.80 |  332494
7964 |     0.51    21.61 | gi|527289996|gb|KE505026.1|       scaffold_38978

#      26     1730 |        4     1722 |     1705     1719 |    99.01 |  332494
```

```
5948 |    0.51    28.90  | gi|527289996|gb|KE505026.1|       scaffold_39157

#    1908    2270 |    2525    2157 |     363     369 |   97.83 |  332494
7964 |    0.11    4.63  | gi|527289996|gb|KE505026.1|       scaffold_38978

#    2743    5779 |    2903    5948 |    3037    3046 |   98.92 |  332494
5948 |    0.91    51.21  | gi|527289996|gb|KE505026.1|       scaffold_39157



#================================================

# Initial stuff

#================================================

$date    = `date`;   # help verbose

# Input control

if ( !$ARGV[0] )

{       # input files basic control

  die "\nUsage: perl parse.coords.cov.pl <filename.coords> [output
name]\n\n";

}

# Output control

$date = `date`;   # help verbose

$in   = $ARGV[0];   # capture input coords file name

if ( $ARGV[1] )

{

  $out = $ARGV[1];   # cmd line output file name

}

else
```

```
{
  $out = $in . "\.parse\.cov\.out";   # output file name, if not in cmd line
}
$log   = $in . "\.parse\.cov\.log";   # log file name


open ( IN,  "$in")   or die "\nCannot open input file $in\n";  # open input file
open ( OUT, ">$out") or die "\nCannot open log file $out\n";   # open output
file
open ( LOG, ">$log") or die "\nCannot open log file $log\n";   # open log
file


$lastref = "0WwW";
$ref     = "1ZzZ";


#================================================
# Reading input file and calculating coverage
#================================================
print "\nPlease, follow log: tail -f $log\n";
print LOG "\n\nStarted at $date\n";
print LOG "\nReading input file and calculating coverage at $date\n";
while ( $linein = <IN> )
{
  chomp $linein;
  if ( $linein =~
```

```
/^\s*\d+\s+\d+\s+\|\s+(\d{1,50})\s+(\d{1,50}).+\..*\|\s+\d+\s+(\d{1,50}).+gi.
*\s+(\w.*)$/ )

  {     # pattern search and capture

    ($ri,$re,$rlen,$ref) = ($1,$2,$3,$4);   # ref_ini, ref_end, ref_lenght and
ref_name

#    print LOG "$rini\,$rend\,$rname\n";   # debug

    if ( $ref ne $lastref )

    {

      ($ri1,$re1) = ($ri,$re);

      $lastref = $ref;

      print LOG "$ref\,$rlen\n";

      print LOG "\,\,\,\,$ref\,$ri1\,$re1\n";

    }

    else

    {

      ($ri2,$re2) = ($ri,$re);

      if ( $ri2 > $re1 and $re2 > $re1 )

      {

        print LOG "\,\,$ri1\,$re1\n";

        ($ri1,$re1) = ($ri2,$re2);

        next;

      }

      else
```

```
    {
      if  ( $ri1 > $ri2 )
      {
        $ri1 = $$ri2;
      }
      if ( $re1 < $re2 )
      {
        $re1 = $re2;
      }
    }
  }
  else
  {
    next;
  }
}

print LOG "\,\,$ri1\,$re1\n";

#================================================
# End things
#================================================
```

```
$date    = `date`;   # help verbose
print LOG "\nFinished at $date\n";


print "Files:
  Input coords: ..... $in
  Output parsed file: .. $out
  Cumulative log: ..... $log


Finished at $date\n";
close LOG;
close IN;
close OUT;
```

APÊNDICE B– Tabela 1 553 indivíduos de 206 subamostras de caiaué (*Elaeis oleifera*) (continua).

REGIÕES	POPULAÇÕES	SUB-REGIÕES	SUBAMOSTRAS[1]	PLANTAS[2]	[1]/ Região	[2]/ Região
REGIÃO DE MANAUS	CALDEIRÃO	REGIÃO NÃO DEFINIDA	7	18	**35**	**95**
	CAREIRO	A. GUTIERREZ	4	9		
		RIO AUTAZ MIRIM	5	14		
		CALDEIRÃO	7	18		
		IGARAPÉ TAPAJÓS	9	27		
	MANACAPURU	FAZ. SÃO JOSÉ	1	3		
	IRANDUBA	TRANCAL	2	6		
RIO AMAZONAS	AMATARI	ALAMBIQUE	3	7	**33**	**90**
		AMATARI	4	12		
		SÃO SEBASTIÃO	4	12		
	AUTAZES	CRIAÇÃO	4	12		
		NOVA ESPERANÇA	4	9		
		QUIRIMIRIM	3	7		
	MAUÉS	SEM POP. DEFINIDA	3	9		
		BOM JARDIM	1	3		
		BOM SOCORRO	1	3		
		ENSEADA	3	8		
		F. SÃO JOAQUIM	3	9		
RIO SOLIMÕES	ANORI	ANORI	2	6	**32**	**92**
		LAGO MIUA	1	3		
	B. CONSTANT	RIO SOLIMÕES	1	3		
	COARI	BARREIRA	1	3		
		IZIDORO	1	3		
		LAGO MAMIÁ GARIBALDE	1	3		
		LAGO MAMIÁ TERRA PRETA	3	9		
		PAXICÁ	9	25		
		PONTA GROSSA	4	11		
	TEFÉ	LAGO CAIAMBÉ	1	3		

		LAGO CATUA	1	3		
		TEFÉ	3	8		
	TONANTINS	SEM POP. DEFINIDA	2	6		
		RUC	1	3		
		TONANTINS	1	3		
RIO NEGRO	ACAJATUBA	ACAJATUBA	1	3	**23**	**63**
		IGARAPÉ DO ARRAIÁ	1	3		
		IGARAPÉ AÇU	8	23		
	BARCELOS	IGARAPÉ CAIAUÉ	2	2		
	MOURA	AIRÃO VELHO	1	3		
		CARVOEIRO	5	14		
		E. CABURIS	3	9		
		NOVO AIRÃO	2	6		
REGIÃO CARACARAI	BR 174	KM 157	1	3	**18**	**53**
		KM 365	4	12		
		KM 490	5	14		
		KM 500	1	3		
		CARACARAÍ	1	3		
	VILA MODERNA	PERIMETRAL NORTE	6	18		
RIO MADEIRA	MANICORÉ	ATININGA	3	7	**65**	**159**
		BACABAL	4	10		
		BARREIRA MUTUPIRI 2	1	2		
		DEMOCRACIA	2	4		
		IGARAPÉ AÇU	4	11		
		ITAPINIMA	3	8		
		LIBERDADE	5	12		
		LAGO SEVERIANO	1	3		
		MANICORÉ	11	27		
		MISSÕES B	9	19		
		REGIÃO NÃO DEFINIDA	10	24		
		RIO MATUPIRI	1	2		
		SANTA HELENA	4	11		
	NOVO ARIPUANÃ	ACARÁ	2	5		
		BALA	1	3		
		PONTA GROSSA	2	6		
		VISTA ALEGRE	2	5		

TOTAL:	**206**	**553**	**206**	**553**

Apêndice C– *Script* em *Perl* desenvolvido para a chamada dos genótipos de marcadores PAVs identificados em caiaué (*Elaeis oleifera*).

```
#!/usr/bin/perl –w

# Script para filtragem do arquivo SAM.

#

# Jaire_hp 18092014

#

# Uso: Perl genotipos.pl <genotipos.csv><frequencia_dart.csv> [nome
output]

#

#================================================

# Inicio do processamento

#================================================

$date = `date`;

# Controle do Input

if ( !$ARGV[1] )

{  #controle básico do input

die "\nUso: perl genotipos.pl <genotipos.csv><frequencia_dart.csv>
[nome output]\n\n";

}

# Controle output

$date = `date`;

$genotipo = $ARGV[0]; #captura a informaÃ§Ã£o de nome do input

$freqdart = $ARGV[1];
```

```
if ( $ARGV[2])

{

  $out = $ARGV[2];

}

else

{

  $out = $genotipo . "\.out"; #nome do output caso nÃ£o especificado na
linha de comando

}

$log = $genotipo . "\.log"; #nome do arquivo de log



open ( GENOTIPO, "$genotipo") or die "\nNão foi possível abrir o
arquivo $genotipo\n"; #abre o arquivo input

open ( FREQDART, "$freqdart") or die "\nNÃo foi possivel abrir o
arquivo $freqdart\n"; #abre o arquivo input

open ( OUT, ">$out") or die "\nNão foi possível abrir o arquivo $out\n";
#abre o arquivo output

open ( LOG, ">$log") or die "\nNão foi possível abrir o arquivo $log\n";
#abre o arquivo de log







#================================================

# Leitura do input e extração das informações

#================================================
```

```
my %HGENOT = ();
my %HDARTFREQ = ();


print LOG "\n\nIniciando em $date\n";
print LOG "\nProcessando informaÃ§Ã£o e extraindo em $date\n";
while ( $linhain = <GENOTIPO> )
{
  chomp $linhain;


if ( $linhain =~ /^([A-Z][0-9]+)\,([a-zA-Z0-9]+)$/ )
{
( $genotnum, $genot ) = ( $1, $2 );
$HGENOT{$genot} = $genotnum;
 }
}


while ( my ($mygenot, $mygenotnum) = each (%HGENOT) )
{
 print OUT "$mygenot\,";
}
while ( my ($mygenot, $mygenotnum) = each (%HGENOT) )
{
 print OUT "$mygenot\,";
```

```
}

while ( $linha = <FREQDART> )
{
chomp $linha;

if ( $linha =~ /^([A-Z]+[0-9]+.[0-9])\,([0-9]+)\,([ACGT]+)\,(\w+)$/ )
{     # busca por padrão e captura
( $cro, $pos, $dart, $map )  = ( $1, $2, $3, $4 );
}
if ( $linha =~ /^([a-zA-Z0-9]+)\,([0-9]+|.*)$/)
{
( $genot2, $freq ) = ( $1, $2 );
    $HDARTFREQ{$dart}{$cro}{$pos}{$map}{$genot2} = $freq;
}
}

for $dart ( keys %HDARTFREQ )
{
  print OUT "\n$dart";
  for $cro ( keys %{ $HDARTFREQ{$dart} } )
  {
    print OUT "\,$cro";
```

```
    for $pos ( keys %{ $HDARTFREQ{$dart}{$cro} } )
    {
     print OUT "\,$pos";
     for $map ( keys %{ $HDARTFREQ{$dart}{$cro}{$pos} } )
     {
      print OUT "\,$map";
      while ( my ($mygenot, $mygenotnum) = each (%HGENOT) )
      {
       if ( defined
$HDARTFREQ{$dart}{$cro}{$pos}{$map}{$mygenot}  )
        {
          print OUT
"\,$HDARTFREQ{$dart}{$cro}{$pos}{$map}{$mygenot}";
        }
        else
        {
          print OUT "\,0";
        }
      }
     }
    }
  }
}
```

```
#================================================

# Finalização

#================================================

$date = `date`;

print LOG "\nFinalizado em $date\n";


print "Arquivos:

    Input CSV: ..... $genotipo

    Output analise:..... $out

    Cumulative log:..... $log


Finalizado em $date\n";


close LOG;

close GENOTIPO;

close FREQDART;

close OUT;
```